P153167.PDF [Page: 3 of 57]

CR834

Managing Uncertainty in Digital Geographical Information: *Survey of Technology*

Phase 2 Final Report

Contract No. W7701-4-1824/01-XSK
SSC File No.XSK94-00100-(303) /A

SUBMITTED TO:

Centre de Recherches Pour La Defense Valcartier BAT 15
2459 Boul Pie XI, NORD
BFC Valcartier
Courcelette, Quebec
G0A 1R0

SUBMITTED BY:

WaterMark Industries Inc.
115 Ascot Drive,
Fredericton, New Brunswick
E3B 6G1
Telephone: 506-454-9189
Facsimile: 506-454-9189
E-mail: watrmark@nbnet.nb.ca

DATE: March 1995

Scientific Authority: Sylvia Lam
Defence Research Establishment Valcartier
2459, Boul. Pie XI, north
Courcelette, Quebec G0A 1R0

Executive Summary

This is the second report of the project entitled "Managing Uncertainty in Digital Geographical Information: Survey of Technology". This report concentrates on the identified issues output from Phase 1 of the project in the light of current uncertainty management techniques, with a specific focus on fuzzy set theoretic methods for representation of uncertainty.

The results of a survey of the state of the technology in uncertainty management techniques are presented, with specific reference to geographical data uncertainty. This survey included reviews of uncertainty data collection, modeling, representation, propagation, visualization - including animation and the application of fuzzy logic in handling spatial data. It is clear from the survey that these are active areas of research, and some of the findings can be implemented in systems to manage uncertainty of spatial data.

This study also identifies areas within the DIGEST format that might be considered to facilitate uncertainty management. The following changes to DIGEST are discussed: resolution of the internal inconsistencies of DIGEST, development of a modeling language for DIGEST, addition of specifications for quality control testing, modification of the lineage data quality representation in DIGEST, enhancement of topological consistency within DIGEST, development of toolkits for generating quality assurance measures and the development of toolkits for viewing uncertainty information. These suggestions for modifications to DIGEST are cross-referenced with the issues outlined in the report of Phase 1 of this project, giving consideration to the need to freeze DIGEST and to provide for a stable development environment.

This report concludes that with these modifications, DIGEST will be more "user friendly" and will be improved in its capability to clearly and precisely specify uncertainty in the geographical data it represents. For DIGEST to become more widely used and accepted, it must overcome the marketplace perception that it is "closed" and difficult to use.

The use of uncertainty management techniques and fuzzy set theory offers some promise of the reduction of risks, due to a lack of certainty of the quality of digital geographical data. Any such risk management technique is attractive to DND and potentially offers a strategic and tactical advantage in operational situations. The progression of managing data uncertainty will be greatly assisted by the construction of uncertainty management toolkits.

© HER MAJESTY THE QUEEN IN RIGHT OF CANADA, AS REPRESENTED BY THE MINISTER OF NATIONAL DEFENSE, 1995.

P153167.PDF [Page: 5 of 57]

Table of Contents

1. INTRODUCTION **4**

2. SURVEY OF UNCERTAINTY MANAGEMENT TECHNIQUES **6**

2.1 GEOGRAPHICAL DATA UNCERTAINTY 8
2.2 COLLECTING DATA ON UNCERTAINTY 11
2.2.1 Direct Error Measurement *11*
2.2.2 Error Deduction *11*
2.3 UNCERTAINTY MODELLING AND REPRESENTATION 13
2.4 UNCERTAINTY PROPAGATION 16
2.5 UNCERTAINTY VISUALIZATION 16
2.6 CURRENT STATUS OF METADATA 18

3. UNCERTAINTY MANAGEMENT AND FUZZY LOGIC **23**

3.1 FUZZY CONTROLLERS 27
3.2 FUZZY PATTERN RECOGNITION 29

4. FUZZY LOGIC AND GEOGRAPHICAL DATA **31**

4.1 EXTENSION PRINCIPLE 36
4.2 EXTENSIONS TO TWO DIMENSIONS 37

5. SUGGESTED MODIFICATIONS TO DIGEST FOR UNCERTAINTY MANAGEMENT **39**

5.1 RESOLVE THE INTERNAL CONSISTENCIES OF DIGEST 39
5.2 MODELLING LANGUAGE EXTENSION 42
5.3 SPECIFICATIONS FOR QUALITY CONTROL TESTING 42
5.4 DATA PROCESSING ERRORS 43
5.5 ENHANCE TOPOLOGICAL CONSISTENCY DESCRIPTION 43
5.6 UNCERTAINTY FOR RASTER DATA 44
5.7 TOOLKITS 44

6. CONCLUSIONS **45**

7. REFERENCES **46**

APPENDIX I. DIGITAL NAUTICAL CHART DATA QUALITY ATTRIBUTES **51**

WaterMark

P153167.PDF [Page: 6 of 57]

List of Figures

Figure 1 - The meaning of six-sigma design [from Smith, 1993]. 7
Figure 2 - Epsilon Definition deterministic form 14
Figure 3 - Epsilon Definition probabilistic form 14
Figure 4 - Epsilon Definition bimodal distribution 14
Figure 5 - Uncertainty model for polygons 15
Figure 6 - Simulation of Line Uncertainty 17
Figure 7 - Fuzzy set characteristic function for the set of apartments close to the University. 23
Figure 8 - Line segment definition for membership functions. 24
Figure 9 - Fuzzy controller output for a lime kiln. 28
Figure 10 - Fuzzy classification results of the butterfly data example [Bezdek, 1981] 29
Figure 11 - Fuzzy nearest-neighbour interpolation example [from Lam, 1994]. 31
Figure 12 - The membership function for fuzzy set "sufficient potassium" [from Kollias and Voliotis, 1991]. 32
Figure 13 - Fuzzy query and result [from Kollias and Voliotis, 1991]. 32
Figure 14 - Maps of the modal fuzzy membership values for each of the three clusters identified for the concept "near" to the Student Center [from Fisher and Orf 1991]. 33
Figure 15 - Boolean (a) and continuous (b) membership functions for class membership [from Heuvelink and Burrough 1993]. 34
Figure 16 - Original soil data from Alberta [from Heuvelink and Burrough 1993]. 34
Figure 17 - Interpolated soil data [from Heuvelink and Burrough 1993]. 35
Figure 18 - Boolean (a) and continuous (b) classification of the soils according to the suitability criteria in equation (15) [from Heuvelink and Burrough 1993]. 36
Figure 19 - Plot of two fuzzy points at locations (2, 6) and (8, 7) [from Xu and Nickerson, 1995]. 37
Figure 20 - DIGEST 40

List of Tables

Table 1 - DIGEST versus Found Issues 4
Table 2 - User Issues 5
Table 3 - Response to cholera vaccine in Chilean children [from Lagos et al, 1995]. 6
Table 4 - Product yield versus product robustness and complexity [from Smith, 1993]. 7
Table 5 - Maximum computed standard deviation in planimetric diagnostic points for map ratings [from Energy, Mines and Resources Canada, 1994]. 8
Table 6 - Crisp set properties [from Klir and Folger, 1988]. 25
Table 7 - Classes of fuzzy set unions and intersections [from Klir and Folger, 1988]. 25
Table 8 - Crop confusion table after relaxation for MSS data [from Di Zenzo et al, 1987]. 30
Table 9 - Cross Reference of Suggested DIGEST Modifications 45

P153167.PDF [Page: 7 of 57]

105-01843/

1. Introduction

Our initial report [WaterMark Industries, 1995] discusses the results of a literature review of uncertainty issues related to geographical data, along with the results of a major user needs survey. DIGEST was reviewed in light of the issues discovered in the literature review and results from the user needs survey. In summary form, these issues are as shown in Table 1.

Table 1 - DIGEST versus Found Issues

Issue		DIGEST Status	Comments
Issue I	Can DIGEST represent these six fundamental components of digital cartographic data quality?	DIGEST has a "source" file quality descriptor intended for this purpose.	Can contain e.g. source datum, datum change constants, grid rotation coefficients
Issue II	Can DIGEST represent both quantitative and nonquantitative secondary method errors?	The "source" file can contain parameters describing scale, translation, and other transformations	No capability for errors due to mislabeling or misclassification
Issue III	Can DIGEST represent statistically meaningful declarations of both quantitative and nonquantitative secondary method errors?	No apparent capability to describe statistical quality tests of the data	Does have attribute accuracy (st. dev. for quantitative and % reliability for qualitative)
Issue IV	Can DIGEST represent data processing errors, and accommodate error propagation among different GIS layers?	"Source" file can contain accurate descriptions of transformations of source material	
Issue V	Can DIGEST represent local uncertainty for raster data?	Can have complete Quality Colour patches (ref. p.A-22) with raster representation. Not intended for general-purpose raster data uncertainty	Need a "cloud cover %" quality indicator for satellite imagery
Issue VI	Can DIGEST accurately represent uncertainty in topological consistency?	No apparent capability here	
Issue VII	Does DIGEST have a set of tools that allow users to view the uncertainty associated with data encoded in it?	None that we could find	

P153167.PDF [Page: 8 of 57]

From the user survey, nine other issues were identified. These are summarized in Table 2 below.

Table 2 - User Issues

User Issue	Description	Related DIGEST Issue(s)
1	Can DIGEST accommodate the Quality Assurance program parameters being adopted by NR Can? This includes process precision, integrity, lineage, accuracy.	I, II, VI
2	How is data uncertainty related to the updating of map layers separately?	I, II, IV
3	What about the concept of re-engineering data attributes for individual datasets?	I, II
4	What about quality issues for the Global Geospatial Information and Services Initiative?	I, II
5	Can DIGEST handle a rule-based system?	IV
6	Can DIGEST address satellite imagery?	V
7	Can DIGEST handle the uncertainty issues relating to the raster versions of the ENC/DNC specification?	V
8	There seems to be a need for some software tools on the quality side.	VII
9	What uncertainty management techniques could be applied to the DIGEST toolbox?	VII

This report looks at these issues in the light of current uncertainty management techniques, with a specific focus on fuzzy set theoretic methods for representation of uncertainty.

WaterMark

2. Survey of Uncertainty Management Techniques

Uncertainty management plays a large role in every scientific and engineering endeavor that relies on measured data. Two important areas unrelated to geographical data processing include medical research and manufacturing processes.

Most medical research that attempts to show the effect of new treatments on humans relies on two groups of patients; the control group and the treatment group. Substantiation of medical hypotheses usually relies on showing that there is a statistically significant difference between the control and treatment groups [e.g. Lagos et al, 1995].

Table 3 shows the primary results from a study of 349 five-to-nine year old Chilean children.

Table 3 - Response to cholera vaccine in Chilean children [from Lagos et al, 1995].

Blood group	N	GMT		Fold rise in titer
		Prevaccination	Postvaccination	
Non-O	67	19	179	9
O	101	21	486	23
P		NS	0.002	0.004

In this study, 178 children received a cholera vaccine, and 171 children received a placebo. GMT stands for geometric mean titer, a measure of the amount of cholera antibodies found in the blood both before and after vaccination. N is the number of children in each blood group. P is the probability of the observed difference between the two blood groups is due to random events (i.e. that the observed difference is not due to the blood type difference). NS stands for not significant. With P = 0.002, this essentially means that we are 99.8% confident that the observed difference in postvaccination rise in antibodies is different for non-O compared to O blood groups. This paper points out the absolutely vital role that statistics plays in managing uncertainty in medical research. Research results are not considered relevant unless they are shown to be statistically meaningful on a large enough sample population. This is true, in general, for all scientific research.

Manufacturing processes depend on statistical analysis for minimizing the number of total defects per unit on manufactured products. Recent improvements in the understanding of how to determine and minimize these defects have substantially improved the overall quality of products manufactured by Motorola [Smith, 1993]. Figure 1 shows the overall meaning associated with six-sigma design.

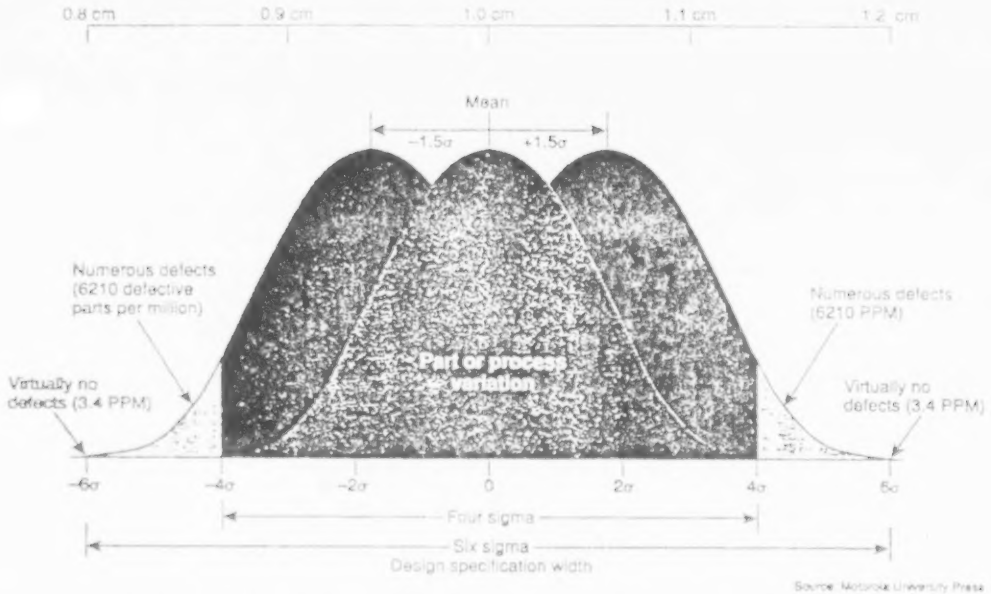


Figure 1 - The meaning of six-sigma design [from Smith, 1993].

In Figure 1, the two bell-shaped curves located at +1.5σ and -1.5σ with respect to the part or process variation bell curve reflect the possible batch-to-batch differences (i.e. shifted results) arising from differences in manufacturing process parameters. With six-sigma design, even with these batch-to-batch differences, the end result is still only a 3.4 part per million (PPM) defect rate. Four-sigma design results in a 6210 PPM defect rate. Table 4 summarizes the importance of the six-sigma design approach to controlling total defects in products containing many parts.

Table 4 - Product yield versus product robustness and complexity [from Smith, 1993].

Complexity, parts/product	Yield, percent			
1	93.54	99.3	99.9	100.0
3	81.84	98.1	99.9	100.0
10	51.27	93.9	99.7	100.0
30	13.48	83.0	99.3	99.9
100	0.13	53.7	97.7	99.9
300	0.00	15.5	93.2	99.9
1000	0.00	0.2	79.2	99.6
3000	0.00	0.0	49.7	98.9
10 000	0.00	0.0	9.7	96.6
30 000	0.00	0.0	0.09	90.3
Robustness				
Std. deviations	±3σ	±4σ	±5σ	±6σ
C_p	1.00	1.33	1.67	2.00
C_{pk}	0.50	0.83	1.17	1.50
PPM	66 810	6210	233	3.4

C_p = capability index, the ratio of the design specification width to the normal variation of the process
C_{pk} = mean shifted capability index PPM defective parts per million

P153167.PDF [Page: 10 of 57]

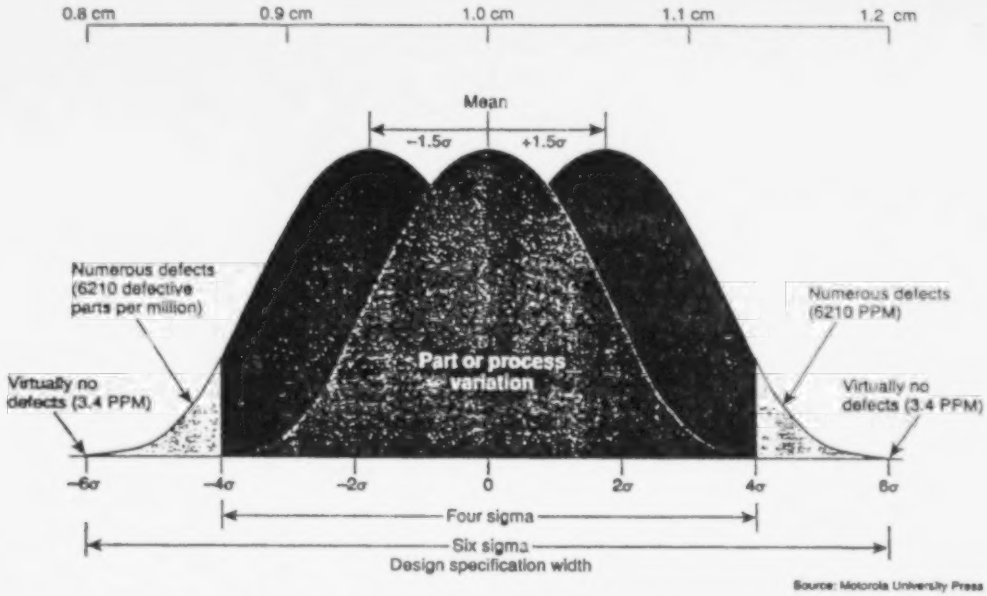


Figure 1 - The meaning of six-sigma design [from Smith, 1993].

In Figure 1, the two bell-shaped curves located at +1.5σ and -1.5σ with respect to the part or process variation bell curve reflect the possible batch-to-batch differences (i.e. shifted results) arising from differences in manufacturing process parameters. With six-sigma design, even with these batch-to-batch differences, the end result is still only a 3.4 part per million (PPM) defect rate. Four-sigma design results in a 6210 PPM defect rate. Table 4 summarizes the importance of the six-sigma design approach to controlling total defects in products containing many parts.

Table 4 - Product yield versus product robustness and complexity [from Smith, 1993].

Complexity, parts/product	Yield, percent			
1	93.54	99.3	99.9	100.0
3	81.84	98.1	99.9	100.0
10	51.27	93.9	99.7	100.0
30	13.48	83.0	99.3	99.9
100	0.13	53.7	97.7	99.9
300	0.00	15.5	93.2	99.9
1000	0.00	0.2	79.2	99.6
3000	0.00	0.0	49.7	98.9
10 000	0.00	0.0	9.7	96.6
30 000	0.00	0.0	0.09	90.3
Robustness				
Std deviations	3σ	4σ	5σ	6σ
C_p	1.00	1.33	1.67	2.00
C_{pk}	0.50	0.83	1.17	1.50
PPM	66 810	6210	233	3.4

C_p = capability index, the ratio of the design specification width to the normal variation of the process.
C_{pk} = mean-shifted capability index. PPM=defective parts per million.

P153167.PDF [Page: 11 of 57]

These two examples point out the important role of statistics in science and engineering. Statistical testing forms a "degree of truth" validation of experimental results [e.g. Walpole and Myers, 1972]. Statistics is based on probability theory [e.g. Papoulis, 1991], which, in turn, relies on axioms regarding the notion of random occurrence of events. Most of what we observe in nature seems to have randomness associated with it at some level.

2.1 Geographical Data Uncertainty

Bartakovitch [1994] outlines the quality controls required for paper map products of the Canada Centre for Mapping. The quality assurance program implemented for National Topographic System (NTS) paper maps is based on the NATO Standardization Agreement [NATO, 1989]. Parameters evaluated include integrity, accuracy, lineage, and process precision. An acceptable quality level is established through stratified random sampling (i.e. by choosing map sheets based on the total number produced in any one batch) and statistical sampling of produced maps. A determination is made of the number of critical, major and minor defects in the map. Map sheets with ≥ 1, 2 or 20 of these defect types, respectively, cause them to be unacceptable.

At least 70 "diagnostic points" are sampled for each map sheet to determine the planimetric accuracy of a map sheet. At least 140 "diagnostic points" are sampled for each map sheet to determine overall relief accuracy for a map sheet. Based on the standard deviation of the difference of these diagnostic points with the actual (ground truth) values for the position and height, the map sheets are rated on a scale as shown in Table 5. A similar rating table exists for height accuracy, with ratings ranging from 0 (best) to 4 (not determined). Complete specifications for Canadian NTS maps are given in Energy, Mines and Resources Canada [1994], including checklists for use by personnel carrying out the quality assurance testing.

Table 5 - Maximum computed standard deviation in planimetric diagnostic points for map ratings [from Energy, Mines and Resources Canada, 1994].

	Scale of map				
Rating	1:25 000	1:50 000	1:100 000	1:200 000	1:250 000
A	12.5m	25m	50m	100m	125m
B	25m	50m	100m	200m	250m
C	50m	100m	200m	400m	500m
D	Poorer than C				
E	Not determined				

Many papers have been written on the subject of geographical data uncertainty. Most models of uncertainty rely on a basic underlying mathematical model of the space in which the data was observed, and subsequently processed or manipulated. True comparisons of geographical data uncertainty with the actual ground truth data are difficult to find, except for some examples of image classification which are compared to the actual ground truth data [e.g. Lee et al, 1987].

P153167.PDF [Page: 12 of 57]

Thapa and Bossler [1992] give a comprehensive overview of geographical data uncertainty. They argue that there are two fundamental methods of data collection; i.e.

1. Primary data is collected directly from the field (e.g. ground surveying, photographs and satellite imagery);
2. Secondary data is collected from existing documents such as maps, charts and graphs.

These two fundamental methods have very different representations of the underlying data uncertainty. In addition, they report that there are six fundamental components of digital cartographic data, namely lineage, positional accuracy, attribute accuracy, logical consistency, completeness and temporal accuracy. Ideally, one would like the data quality representation to somehow encompass all of these components.

Some examples of errors occurring in secondary methods of data collection that don't occur in the primary methods include eleven quantitative errors; i.e.

(1) error in plotting control points,
(2) compilation error,
(3) error introduced in drawing,
(4) error due to generalization,
(5) error in map reproduction,
(6) error in colour registration,
(7) error introduced by the deformation of the material,
(8) error introduced by the use of uniform scale,
(9) feature definition errors,
(10) error introduced due to feature exaggeration, and
(11) errors in digitization.

In addition three nonquantitative secondary method errors include

(1) error due to mislabeling,
(2) errors due to misclassification (of e.g. roads, symbols, woodlands), and
(3) errors due to feature coding (e.g. contour labelled improperly).

These errors are not normally detected by any standard uncertainty measurement methods.

Accuracy standards exist which state that, e.g. for horizontal accuracy "for maps on publication scales larger than 1:20,000, not more than 10 percent of the points tested shall be in error by more than 1/30 inch, measured on the publication scale; ...". Thapa and Bossler argue that while standards exist for primary data collection methods, no such standards exist for secondary data collection methods and are sorely needed. Amrhein and Griffith [1991] present one technique for a mean square error measure of a digitized line. Bolstand et al [1990] also discuss the statistical characterization of positional error in manually digitized map data.

Hunter [1992] suggests that there are two forms of error in spatial databases; positional error and attribute error. These error forms are due to three error sources; i.e. data collection and manipulation, data processing and data misuse. Data processing errors are often overlooked, and

P153167.PDF [Page: 13 of 57]

should be able to be accounted for in the overall data quality measure. This processing occurs for both vector and raster data types [e.g. Veregin, 1989]. In addition to the processing errors arising when the data is generated, there are errors arising from the combination of different data layers within a GIS [Lanter and Veregin, 1992].

Goodchild [1992] argues that data error and uncertainty is treated more appropriately from the perspective of fields (representing regions or segments) rather than individual objects (e.g. points, lines and areas). Objects might be stripped of information on which a useful model of error might be based, by e.g. low-pass boundary filtering or small region deletion. Goodchild and Min-hua [1988] point out that class membership probabilities contain useful information on source error, which should probably be maintained by the GIS.

Edwards [1992] states that a GIS should be able to manage local uncertainty as opposed to representing only summary uncertainty for entire geographical datasets. His work looks at integrating image analysis systems with GIS, and he discusses a novel approach to generating "uncertainty maps" of raster data.

Uncertainty for geographical information arises from a model of observations of the real world. An observation vector l is assumed to exist. This observation vector is normally derived from some time sequence of observations in space, and can contain three components, as follows [e.g. Vanicek and Krakiwsky, 1982; Mikhail, 1976]

$$l(t) = \hat{\ell}(t) + p(t) - v(t) - s(t) \qquad (1)$$

where $\hat{\ell}(t)$ = expected value of the observable,
$p(t)$ = systematic component,
$v(t)$ = the statistically independent residual ($E(v) = 0$), and
$s(t)$ = the statistically dependent residual.

To determine the desired set of parameters $\mathbf{x}$, a mathematical model of the form

$$\mathbf{f}(\mathbf{x}, l) = \mathbf{0} \qquad (2)$$

is solved for $\mathbf{x}$. For nonlinear mathematical models relating the observables to the parameters, an iterative solution to the parameters results in an estimate of the parameters $\hat{\mathbf{x}}$, an estimate of the observables $\hat{l} = l + \hat{\mathbf{r}}$, and an estimate of their covariance matrices $\mathbf{C}_{\hat{\mathbf{x}}}$ and $\mathbf{C}_{\hat{\mathbf{r}}}$. These covariance matrices represent the uncertainty associated with geographical data estimated in this way. Commonly, a level of statistical significance greater than one standard deviation is achieved by multiplying the covariance matrices by a constant k (e.g. k = 2.146 to achieve the 90% confidence interval associated with parameters in two dimensions).

A particular variant of interpolation known as Kriging (first used in the mining industry to interpolate ore densities) is also used for "gridded" interpolation of geographical data [e.g.

P153167.PDF [Page: 14 of 57]

Heuvelink et al, 1989]. Kriging estimation can also determine the estimated uncertainty of the interpolated data points.

2.2 Collecting Data on Uncertainty

There are two methods of collecting data on uncertainty. One is by direct measurement and the other is by deduction. Direct measurement means the use of more accurate reference data to directly measure the amount of uncertainty. Deduction means the derivation, using analytical or statistical methods, of uncertainty measurements for a process or a product.

2.2.1 Direct Error Measurement

Direct measurement of uncertainty of spatial data is also called ground truthing or comparison to the source. But absolute truth cannot be attained [Parratt, 1971]. Therefore the truth or reference often means an independent source of higher accuracy. Higher accuracy usually implies that data are of a larger scale, more recent, and more carefully measured with more accurate measuring instruments. Higher accuracy also means less processed data. This direct measurement of uncertainty could be for positional uncertainty, and thematic uncertainty as well.

Direct measurement is time consuming, but it is a useful technique for generating uncertainty measurements at selected points so that deduction methods can be used to estimate the uncertainty at other locations.

2.2.2 Error Deduction

When direct error measurement is not possible or too expensive, analytical methods to deduce the magnitudes of error must be used. The estimation can be applied either to a process or a product.

When we deduce the uncertainty measurement of a product, we estimate its degree of uncertainty without having to perform a direct measurement on every single piece of data. This method is also called internal evidence [Antenucci, et al., 1991].

There are different ways by which uncertainty estimations can be achieved. One method is to identify the steps involved in making the product, create an error budget for each step, and obtain an estimation from the error budget. For example, the error of aerotriangulation can be deduced by identifying the error sources and estimating the magnitude of error in each of the sources. For aerotriangulation, the error sources would include film , camera, stereoplotting instrument, instrument operator, earth curvature, and map projection [Slama, 1980]. Thapa and Bossler [1992] have also investigated the error sources of mapping and have illustrated the technique of combining entries from an error budget for mapping.

P153167.PDF [Page: 15 of 57]

The interaction of the different error sources could be complex, and the modeliing of the combined result is not always trivial. Amrhein [1991] has tackled the problem of modelling the combined effect of line digitizing and line generalization errors.

When we deduce the uncertainty factor of a process, like digitizing, we try to determine the average amount of error that could be generated by that process. If this is performed on an instrument, it is called calibration. Bolstad et al. [1990] describes a method of estimating manual digitizing errors, and Derenyi et al. [1993] describes a method of estimating semi-automatic digitizing errors.

In order to determine the error factor of a process, a trial measurement method is usually adopted which often produces a set of observations or measurements. The meaning of the term 'trial' implies that the set of observations or measurements should be independent of each other and is carried out under identical conditions. Because of random errors, trial observations tend to cluster about some more or less central value, and the clustering shows a type of statistical regularity that appears in the shape of what is generally called the frequency distribution. Therefore we can deduce the uncertainty factor from the limited number of trial measurements. Usually this is done by means of statistical tools.

Another method is by interpolation. Methods such as Kriging [Oliver and Webster, 1990] are convenient ways of obtaining these information given the uncertainty measurement at certain locations. Image classification techniques also generate estimates of uncertainty as a by-product of processing the data [Goodchild, 1988].

Simulation is a technique that can be combined with other methods to estimate uncertainty. Veregin [1994] describes a simulation technique for estimating error propagation in buffer operations. The technique involves a buffer layer with simulated errors.

We have discussed earlier that Kriging is a popular uncertainty deduction technique. However it is only used to estimate the uncertainty of attribute data at a location. The attribute can be height measurement at a point or the population at a certain location. There is no provision in the Kriging model to include uncertainty in the planimetric locations. That makes it rather difficult to estimate the amount of uncertainty in elevation due to uncertainties in the planimetric position of the control points. The method of simulation [Zhou, 1995] combined with any interpolation method can be used to accommodate this additional uncertainty.

The proposed simulation method works like this. From the distribution of positional error at each of the known control points, we randomly generate a set of control points from their distributions. For each set we will perform an interpolation, say by using bi-cubic splines. By repeatedly performing the same interpolation using multiple sets of control points, an estimation of the uncertainty in elevation can be obtained. The technique is general enough to allow estimation of other parameters of uncertainty.

WaterMark

P153167.PDF [Page: 16 of 57]

2.3 Uncertainty modelling and representation

There are two methods of partitioning space for data modelling; *image-based and object-based.* An image-based partition first subdivides space into regular cells and then assigns attributes to each cell to indicate what spatial feature it belongs to. This results in what we commonly called a raster representation and the cells are called pixels. An object-based partition first determines the outlines of objects, subdivides space by those outlines, and then assigns attributes to each object. This results in what we commonly called a vector representation and the objects are sometimes called features.

Uncertainty modelling can also be classified as image based or object based. Image based uncertainty modelling is used more widely, and it assigns a measurement of uncertainty to each pixel. The uncertainty measurements can indicate the positional or, more commonly, the thematic uncertainties of the object occupying that pixel. When this method is used to indicate the thematic uncertainty of the objects, a classification of the pixels is first obtained. Let us suppose that this results in n classes called C_1 to C_n. Because the class boundaries are fuzzy, a pixel could belong to several classes with different degrees of probability. Suppose we can determine, by using one of the uncertainty measurement or deduction methods mentioned in the previous section, the probability of each pixel belonging to each of the n classes. We can then create a probability vector (P_1, P_2, ... P_n) for each pixel where P_i is the probability of that pixel belonging to C_i.

In an object-based model, uncertainty measurements are associated with spatial objects which are usually represented as points, lines, and polygons. Each spatial object has a positional and a thematic component. Its positional component is based on the coordinates of the points or lines making up the object. Polygons are usually composed of lines and lines are composed of points. Its thematic component is based on other attributes.

The most basic representation of positional uncertainty is that for a point. For a two-dimensional point, the x and y coordinates make up the positional component and all other attributes make up the thematic component. That means the z coordinate of a point can be regarded as a thematic component for a two-dimensional point or we can treat the point as a three-dimensional object and the z coordinate is part of the positional component. The choice depends on the general dimensionality of the database. In a conventional map, most line features are two dimensional but some point features, such as spot heights and soundings, are three dimensional. In that case, it might be more convenient to treat the occasional z as a thematic attribute.
A popular model for point positional uncertainty is to use an error ellipse around a point. A line is composed of points, and hence line positional uncertainty could be reduced to the uncertainty of the component points. The combined effect is an uncertainty band, called an epsilon band, around the line.

There are several methods of defining an epsilon band [Dunn et al., 1990]. In the deterministic form, the band with a width of epsilon is created on both sides of the line (Figure). The distribution function provides no measurement of probability in that it only shows the region (epsilon in width) in which the true line can fall.

WaterMark

P153167.PDF [Page: 17 of 57]

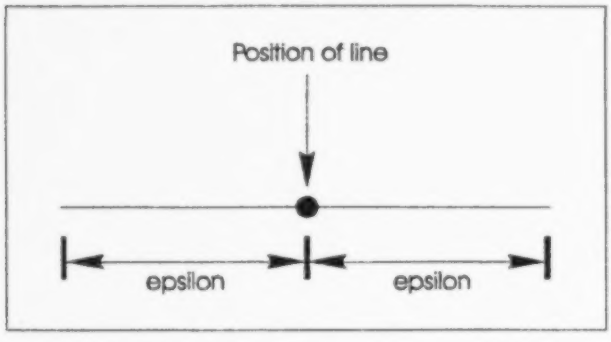


Figure 2 - Epsilon Definition deterministic form

In the probabilistic form, a band is associated with a probability function to indicate the likelihood of the line deviating from its depicted position. (Figure 3)

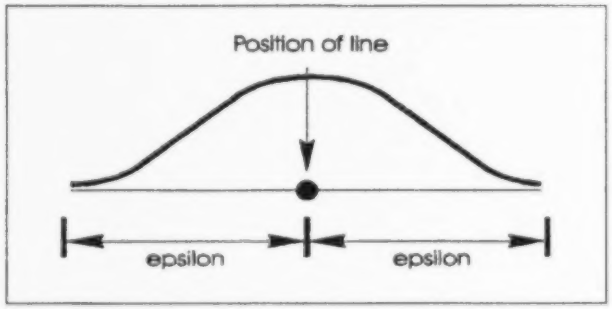


Figure 3 - Epsilon Definition probabilistic form

A more elaborate model is the bimodal distribution, which assumes that the position of the true line is consistently on one side of the depicted line. Hence there are two distribution functions involved, one on each side of the line. (Figure 4)

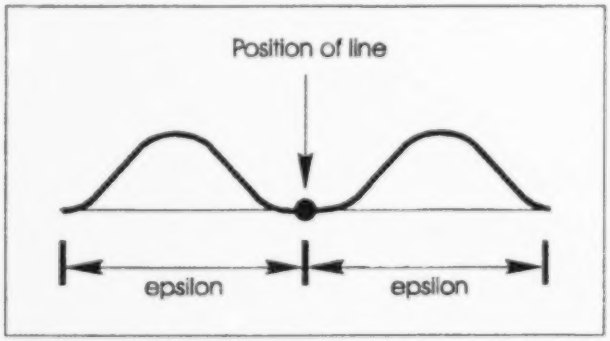


Figure 4 - Epsilon Definition bimodal distribution

P153167.PDF [Page: 18 of 57]

A polygon encloses a classified object and is delineated by an outline. A fuzzy outline can be used to represent either positional or thematic uncertainty because they both affect the location of the outline. A convenient method of representing such a fuzzy outline is to use one of the epsilon methods mentioned above.

Zhou [1995] argued that this approach provides a method of reducing the storage space required. He suggested that an uncertainty model for polygons can be based on a core surrounded by a series of boundaries with homogenous fuzziness (Figure 5).

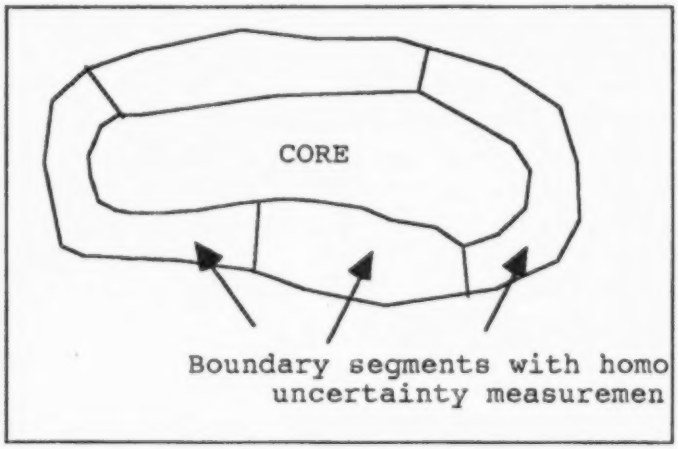


Figure 5 - Uncertainty model for polygons

P153167.PDF [Page: 19 of 57]

2.4 Uncertainty Propagation

Error propagation is the migration of uncertainty from one operation to another, resulting in an increase in the magnitude of error. Hence uncertainty propagation modelling and monitoring are essential components of uncertainty management. There is, however, currently lack of a standard framework for modelling error propagation through sequences of data transformation functions [Lanter and Veregin, 1992].

A spatial operation might or might not produce new objects. The intersection of two polygons will generate new polygons, the intersection of two lines will produce a point, and so on. To monitor the propagation of uncertainty under these operations, two techniques could be used. One is to model the effect of the operator on uncertainty, analyze the final uncertainty from uncertainty measurements of the original data, and then store the resulting uncertainty for input to the next step of the operation.

Another method is to not store the final uncertainty as a single value for the resulting object but to store the original uncertainty measurements and the operator involved. For example, if two polygons A and B from two layers have uncertainty measurements E(A) and E(B), and an overlay operation creates a new polygon C, then the original values of E(A) and E(B) and the logical operator used to create the result should be all captured in polygon C.

The first method is more compact in storage while the second method would keep a history of the operators applied on the object so far. If the second method is used, then an error propagation model must be applied later to derive an uncertainty measurement. Because the model is applied only when a measurement is required, we can apply the best model available at the time.

If an operation does not create new objects, then error propagation analysis must be done immediately and a result generated for the use of the next step in the operation. Overlay of map layers in either vector or raster form always produce new objects. In the raster case, the new object created is a new raster layer.

In any case, there is a need to model the error propagation of an operation. For logical operators, this is not a serious problem. For the more complicated operators, some special technique would have to be used to find the model. Veregin [1994], for example, has addressed the problem of error propagation for the buffering operation. Other issues related to error propagation modelling have been discussed by Heuvelink et al. [1989] and Heuvelink and Burrough [1993]. In particular Heuvelink et al. [1989] described a prototype software package for error propagation in the raster domain. The problem of error propagation for vector data has received less attention.

2.5 Uncertainty Visualization

Uncertainty visualization is a challenging cartographic problem. The main objective here is to graphically depict the uncertainty of data, and the challenge is to make optimal use of the limited

P153167.PDF [Page: 20 of 57]

space normally available on the display surface. A survey of recent visualization techniques is given in Hunter and Goodchild [1993]

Uncertainty visualization techniques depend very much on the data representation and error model used. A natural choice is to use circles and ellipses to represent point uncertainty, and to use bands to represent line uncertainty. For area-based uncertainty visualization, an uncertainty surface is commonly used. Variations of these techniques include the use of the third-dimension, colours of different intensity, patterns of different visual values, and even animation to show the fuzziness of uncertain data.

Animated methods of visualization are proposed by Fisher [1992] to show the uncertain nature of dot maps, soil maps, and classified satellite imagery. A traditional dot map shows the dots each representing, for example, a count of the population at that location. This is a static display conveying the idea that the count applies to that exact location. In reality, the dot location is not exact but is usually placed arbitrarily within a region from which the statistics (such as a census tract) were collected. To show the uncertain location of the dots, their locations are randomly moved with their regions in real-time.

Soil maps are normally shown as choropleth maps with a uniform pattern applied to each zone. In reality, however, the zones could also belong partially to other classes since soil boundaries are not exact. To show this sampling uncertainty, the classification of the zones are randomly varied from one to another in real-time.

For classified satellite imagery, the likelihood of a pixel belonging to a certain class is known. A similar real-time class switching technique can be applied, but the number of times a pixel stays in a class can be controlled by its likelihood of being in that class.

A dynamic technique for showing the uncertainty of lines can be developed in a similar way. Dutton [1992] simulated line segments by connecting uncertain endpoints. In the model, each point is assigned a circular locus uncertainty that represents, for example, how far a point may wander from the centre of a stream and still be in the water. There are a multitude of line segments that might be generated to connect adjacent points (see Figure 6). In a dynamic display technique, the different connecting lines can be chosen at random, thus creating the perception of a fuzzy line.

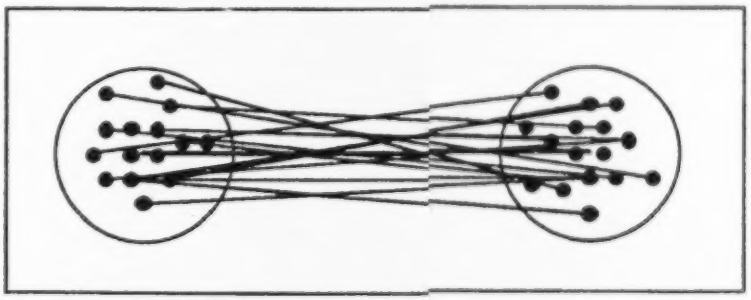


Figure 6 - Simulation of Line Uncertainty

2.6 Current Status of Metadata

The Spatial Data Transfer Standard (SDTS) [USGS, 1994a, 1994b] serves as a national spatial data transfer mechanism for the United States. This standard is a culmination of more than 10 years work by a large number of groups, including the USGS SDTS Technical Review Board, the USGS Digital Cartographic Data Standards Task Force, the Federal Interagency Coordinate Committee for Digital Cartography (Standards Working Group), and the American Congress on Surveying and Mapping (National Committee for Digital Cartographic Data Standards). Besides being a comprehensive solution to the transfer of spatial data, this standard also includes a **required** data quality report for each dataset transmitted using the standard. There are five portions to the data quality report, including lineage, positional accuracy, attribute accuracy, logical consistency, and completeness.

The following list has been complied by putting all data quality descriptors that we have collected from standards, together. This would provide a complete list of data qualifiers at this stage of technology. We have used the FGDC Contents Standards of Digital Geospatial Metadata as the framework to build the list. An indicator of A,B, or C in front of an item indicates if it is part of the data quality descriptors of DIGEST A, B, or C respectively. A description of the three encapsulations of DIGEST is given later.

Product Specification -- a specification for the product-specific standard used to produce this dataset.

(A,B,C) **Name of Product Specification**

(A,B,C) **Date of Product Specification**

(A,B) **Specification Amendment**

Attribute Accuracy -- an assessment of the accuracy of the identification of entities and assignment of attribute values in the data set.

Attribute Accuracy Report -- an explanation of the accuracy of the identification of the entities and assignments of values in the data set and a description of the tests used.

Quantitative Attribute Accuracy Assessment -- a value assigned to summarize the accuracy of the identification of the entities and assignments of values in the data set and the identification of the test that yielded the value.

(A,B,C) **Attribute Accuracy Value** -- an estimate of the accuracy of the identification of the entities and assignments of attribute values in the data set.

Attribute Accuracy Explanation -- the identification of the test that yielded the Attribute Accuracy Value.

P153167.PDF [Page: 22 of 57]

(A,B,C) **Collection Specification** -- Name of specification used to collect data.

(A,B,C) **Logical Consistency** -- an explanation of the fidelity of the relationships in the data set and the tests used.

Clipping indicates if an object has been truncated for this transfer.

Topological consistency

Node identifies if all intersection of geometric objects occur only at explicit positions defined by other geometric objects.

Overlap identifies if objects overlap.

Share identifies if common objects are defined only once.

Entity identifies if all entities are represented directly by a geometrical entity, for example one polygon per lake and not many lines making up the outline of a lake.

Valid Geometry and **relationships** identifies if the geometric entities have valid geometry and relationship with others.

Topographic consistency tests identifies if features are consistent with respect to topography. For instance, that all rivers run downhill, all lake outlines and coastlines are level, and no contours intersect.

Other consistency tests contain a list of integrity test information including the following:

Test name identifies the name of the test.

Submission date

Completior date

Passed indicates if the data set passed the test.

Details contains results of the test either in the form of quantitative measurements or a detail check list.

Software identifies the software used for the test.

Testing authority

WaterMark

(A,B,C) **Completeness Report** -- information about omissions, selection criteria, generalization, definitions used, and other rules used to derive the data set.

Positional Accuracy -- an assessment of the accuracy of the positions of spatial objects.

Horizontal Positional Accuracy -- an estimate of accuracy of the horizontal positions of the spatial objects.

Higher order source identifies the higher order source using in the comparison.

Horizontal Positional Accuracy Report -- an explanation of the accuracy of the horizontal coordinate measurements and a description of the tests used.

Quantitative Horizontal Positional Accuracy Assessment -- numeric value assigned to summarize the accuracy of the horizontal coordinate measurements and the identification of the test that yielded the value.

(A,B,C) **Horizontal Positional Accuracy Value** -- an estimate of the absolute and relative accuracies of the horizontal coordinate measurements in the data set expressed in (ground) meters.

Horizontal Positional Accuracy Explanation -- the identification of the test that yielded the Horizontal Positional Accuracy Value.

Vertical Positional Accuracy -- an estimate of accuracy of the vertical positions in the data set.

(A,B,C) **Vertical Positional Accuracy Report** -- an explanation of the absolute and relative accuracies of the vertical coordinate measurements and a description of the tests used.

Quantitative Vertical Positional Accuracy Assessment -- numeric value assigned to summarize the accuracy of vertical coordinate measurements and the identification of the test that yielded the value.

Vertical Positional Accuracy Value -- an estimate of the accuracy of the vertical coordinate measurement in the data set expressed in (ground) meters.

Vertical Positional Accuracy Explanation -- the identification of the test that yielded the Vertical Positional Accuracy Value.

Temporal Accuracy -- likely time error.

Control

P153167.PDF [Page: 24 of 57]

Horizontal and vertical controls describing the

Authority is the agency responsible for the control,

Adjustment order (1st, 2nd, 3rd, etc.),

Integration status indicates whether the survey is tied to other networks.

Collection agency is the agency responsible for data collection.
Final collection date.

Adjustment agency.

Final adjustment date.

Currency

(A,B,C) **Edition number of data set**

(A,B,C) **Date of creation**

(A,B,C) **Date of revision**

(A,B) **Recompilation count**

(A,B) **Revision count**

Revision cycle

Date for next revision

Lineage -- information about the events, parameters, and source data which constructed the data set, and information about the responsible parties.

Source Information -- list of sources and a short discussion of the information contributed by each.

(C) **Name of included source file**

Source Citation -- reference for a source data set.

WaterMark

P153167.PDF [Page: 25 of 57]

Source Scale Denominator -- the denominator of the representative fraction on a map (for example, on a 1:24,000-scale map, the Source Scale Denominator is 24000).

Type of Source Media -- the medium of the source data set.

Source Time Period of Content -- time period(s) for which the source data set corresponds to the ground.

(A,B,C) **Source Currentness Reference** -- the basis on which the source time period of content information of the source data set is determined.

Source Citation Abbreviation -- short-form alias for the source citation.

Source Contribution -- brief statement identifying the information contributed by the source to the data set.

Process Step -- information about a single event.

Process Description -- an explanation of the event and related parameters or tolerances.

Source Used Citation Abbreviation -- the Source Citation Abbreviation of a data set used in the processing step.

Process Date -- the date when the event was completed.

Process Time -- the time when the event was completed.

Source Produced Citation Abbreviation -- the Source Citation Abbreviation of an intermediate data set that (1) is significant in the opinion of the data producer, (2) is generated in the processing step, and (3) is used in later processing steps.

Process Contact -- the party responsible for the processing step information.

Cloud Cover -- area of a data set obstructed by clouds, expressed as a percentage of the spatial extent.

WaterMark

P153167.PDF [Page: 26 of 57]

3. Uncertainty Management and Fuzzy Logic

Fuzzy logic arises from the notion of fuzzy sets originally introduced by Zadeh [1965]. Substantial research and investigation into the application of fuzzy sets to a wide variety of domains has been performed since the original papers appeared. Complete texts [e.g. Dubois and Prade, 1980; Klir and Folger, 1988] have been written to present the theory and applications in a comprehensive light. Despite this, the theory still has its opponents [Elkan, 1994] who provide critical observations which will probably lead to improvements in the theory and the way it is applied [e.g. Yager, 1994; Zadeh, 1994].

A fuzzy set extends set theory to include the notion of partial membership. A fuzzy set is defined by its membership function. For example, the fuzzy set A describing the set of "apartments close to the University" can be defined by the membership function shown in Figure 7.

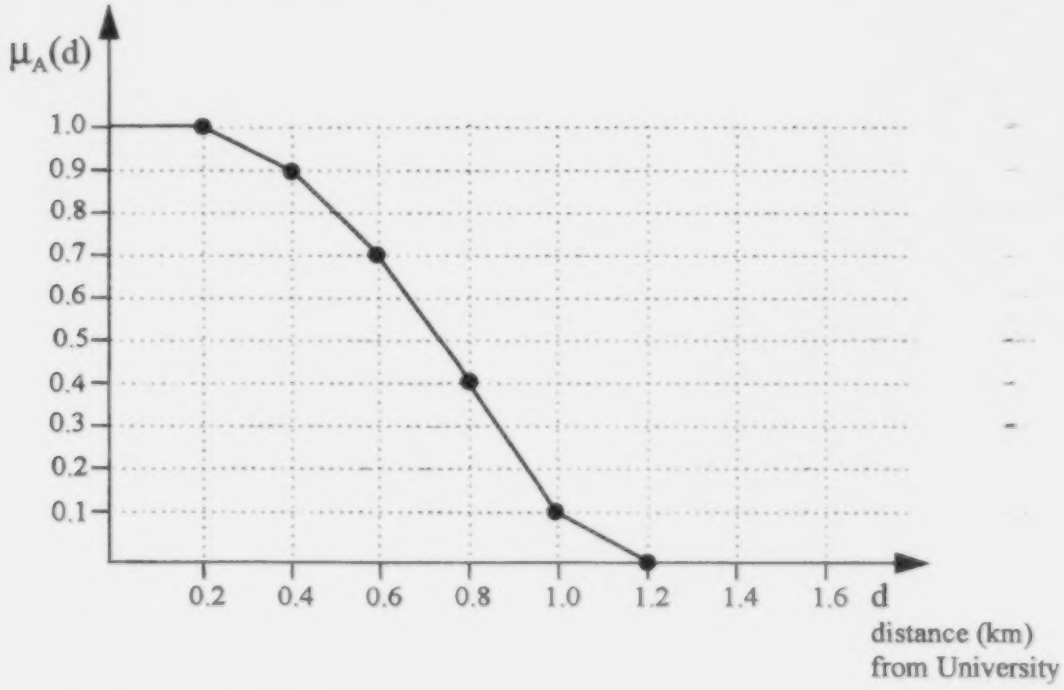


Figure 7 - Fuzzy set characteristic function for the set of apartments close to the University.

This figure illustrates that membership in the set "close to the University" can be partial, i.e. between 0 and 1 in the range of 0.2 to 1.2 km. For example, an apartment located at 0.7 km distance from the University has a membership value of 0.7 in the fuzzy set "close to the University".

A general-purpose function for defining membership functions is shown in Figure 8.

$$L(x;a,b,c,d) = \begin{cases} 0 & x < a-c \\ \dfrac{x-(a-c)}{c} & a-c \le x \le a \\ 1 & a < x < b \\ 1-\dfrac{x-b}{d} & b \le x \le b+d \\ 0 & b+d < x \end{cases}$$

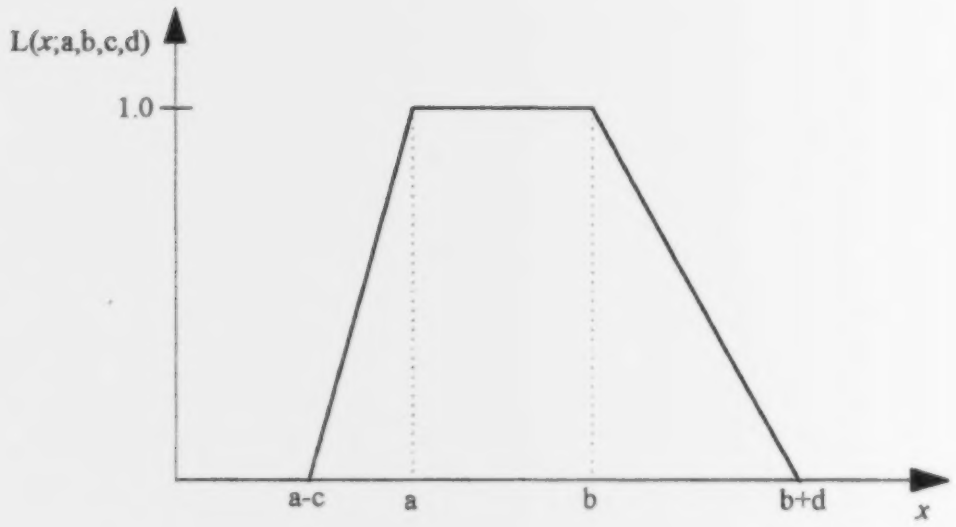


Figure 8 - Line segment definition for membership functions.

This line segment definition is very fast to evaluate, and can accommodate a wide variety of membership function shapes.

Table 6 lists the properties of crisp sets.

P153167.PDF [Page: 28 of 57]

Table 6 - Crisp set properties [from Klir and Folger, 1988].

Involution	$\overline{\overline{A}} = A$
Commutativity	$A \cup B = B \cup A$ $A \cap B = B \cap A$
Associativity	$(A \cup B) \cup C = A \cup (B \cup C)$ $(A \cap B) \cap C = A \cap (B \cap C)$
Distributivity	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$ $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
Idempotence	$A \cup A = A$ $A \cap A = A$
Absorption	$A \cup (A \cap B) = A$ $A \cap (A \cup B) = A$
Absorption of complement	$A \cup (\overline{A} \cap B) = A \cup B$ $A \cap (\overline{A} \cup B) = A \cap B$
Absorption by X and $\varnothing$	$A \cup X = X$ $A \cap \varnothing = \varnothing$
Identity	$A \cup \varnothing = A$ $A \cap X = A$
Law of contradiction	$A \cap \overline{A} = \varnothing$
Law of excluded middle	$A \cup \overline{A} = X$
DeMorgan's laws	$\overline{A \cap B} = \overline{A} \cup \overline{B}$ $\overline{A \cup B} = \overline{A} \cap \overline{B}$

Fuzzy sets commonly lose the "law of excluded middle" and "law of contradiction" properties. This is due to the definition of the intersection and union operations in the following manner:

$$i(a,b) = \min(a, b) \quad (3)$$
$$u(a,b) = \max(a, b) \quad (4)$$

This allows the computation of the intersection of two fuzzy sets as follows:

$$\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x)) \quad (5)$$

As shown in Table 7, many other definitions of fuzzy set union and intersection have been developed over the years.

Table 7 - Classes of fuzzy set unions and intersections [from Klir and Folger, 1988].

Reference	Fuzzy Unions	Fuzzy Intersections	Range of Parameter
Schweizer & Sklar [1961]	$1 - \max[0, (1-a)^{-p} + (1-b)^{-p} - 1)]^{1/p}$	$\max(0, a^{-p} + b^{-p} - 1)^{-1/p}$	$p \in (-\infty, \infty)$
Hamacher [1978]	$\frac{a + b - (2-\gamma)ab}{1 - (1-\gamma)ab}$	$\frac{ab}{\gamma + (1-\gamma)(a + b - ab)}$	$\gamma \in (0, \infty)$
Frank [1979]	$1 - \log_s\left[1 + \frac{(s^{1-a} - 1)(s^{1-b} - 1)}{s - 1}\right]$	$\log_s\left[1 + \frac{(s^a - 1)(s^b - 1)}{s - 1}\right]$	$s \in (0, \infty)$
Yager [1980]	$\min[1, (a^w + b^w)^{1/w}]$	$1 - \min[1, (1-a)^w + (1-b)^w)^{1/w}]$	$w \in (0, \infty)$
Dubois & Prade [1980]	$\frac{a + b - ab - \min(a, b, 1-\alpha)}{\max(1-a, 1-b, \alpha)}$	$\frac{ab}{\max(a,b,\alpha)}$	$\alpha \in (0, 1)$
Dombi [1982]	$\frac{1}{1 + \left[\left(\frac{1}{a} - 1\right)^{-\lambda} + \left(\frac{1}{b} - 1\right)^{-\lambda}\right]^{-1/\lambda}}$	$\frac{1}{1 + \left[\left(\frac{1}{a} - 1\right)^{\lambda} + \left(\frac{1}{b} - 1\right)^{\lambda}\right]^{1/\lambda}}$	$\lambda \in (0, \infty)$

P153167.PDF [Page: 29 of 57]

Using a different definition for intersection and union operations changes the way in which fuzzy logic works. By far the most commonly used definitions are the original Zadeh definitions shown in equations (3) and (4).

Fuzzy logic is used when reasoning about objects described using fuzzy sets [e.g. Martin-Clouaire, 1989]. This allows one to draw inferences about objects using rules of the form

$$\begin{array}{l} \text{If X is A then Y is C} \\ \quad \text{X is A'} \\ \hline \qquad\qquad \text{Y is C'} \end{array} \qquad (6)$$

where A, C, A' and C' are all fuzzy sets. The fuzzy set representation for C' is computed as

$$C' = A' \circ (A \rightarrow C) \qquad (7)$$

where $\circ$ means composition, and $A \rightarrow C$ means "A implies C".

Fuzzy set A composed with relation R is commonly defined as

$$A \circ R = \bigvee_x (\mu_A(x) \wedge \mu_R(x,y)) \qquad (8)$$

for $x \in X$, $y \in Y$, R defined on $X \times Y$. The symbol $\bigvee_x$ means "logical or over x", or, with the standard Zadeh definitions of intersection and union, the union or "maximum over all x". The symbol $\wedge$ means "logical and" or intersection. In set theory, the above expression would normally be written as

$$A \circ R = \sup_x \min(\mu_A(x), \mu_R(x,y)) \qquad (9)$$

where sup stands for supremum, meaning "least upper bound".

As a simple example, consider the fuzzy set "small" composed with a relation R defined as "approximately equal to". We have

$$\text{small} = 1/1 + 0.6/2 + 0.2/3 + 0/4,\ X = Y = \{1, 2, 3, 4\} \qquad (10)$$

The relation R = "x approximately equal to y" is defined as

R y / x	1	2	3	4
1	1	0.5	0	0
2	0.5	1	0.5	0
3	0	0.5	1	0.5
4	0	0	0.5	1

WaterMark

P153167.PDF [Page: 30 of 57]

and

$$\text{small} \bullet \text{approximately equal} = 1/1 + 0.6/2 + 0.5/3 + 0.2/4 \qquad (11)$$

or "approximately small".

Two application areas where fuzzy logic has demonstrated its power are in the areas of fuzzy controllers [e.g. Sugeno, 1985] and in pattern recognition [e.g. Bezdek, 1981].

3.1 Fuzzy Controllers

Fuzzy controllers use natural language-like rules to specify controller output actions given input variable descriptions. For example, in the control of a lime kiln, one of the control rules reads as follows:

IF the change in kiln drive torque is NEGATIVE
AND the free lime content is LOW
THEN adjust the fuel by a SMALL POSITIVE amount

These rules can be easily understood by humans, and modified as required. The rule terms are specified by fuzzy set definitions. Multiple rules are combined using the rule of compositional modus ponens [e.g. Lee, 1990], which normally is carried out via a max-product composition. With n rules, the max-product composition for fuzzy controllers yields

$$\mu_{C'} = \bigcup_{i=1}^{n} \alpha_i \mu_{C_i} \qquad (12)$$

where i stands for rule i, $\mu_{C'}$ is the resultant characteristic function of the output, μ_{C_i} is the characteristic function of the output linguistic variable for rule i,

$$\alpha_i = \bigcap_{j=1}^{m} x_j \qquad (13)$$

for $x_j = \mu_{A_j}(a_j)$; i.e. the input linguistic variable A_j characteristic function evaluated at the single input value a_j. There are m input variables, $\cap$ stands for intersection (i.e. minimum) and $\cup$ stands for union (i.e. maximum).

Figure 9 shows a set of rules for the fuzzy controller of a lime kiln. The example shows nine rules of the type given at the beginning of section 3.1. The example also shows the resultant output fuzzy set for the resulting adjustment of fuel rate given the two input values "change in kiln drive torque = -1.2%/h" and "free lime content = 0.54%". A real number for control of the plant (in this case -0.048m^3/h) is determined through the defuzzification process [e.g. Cox, 1994].

P153167.PDF [Page: 31 of 57]

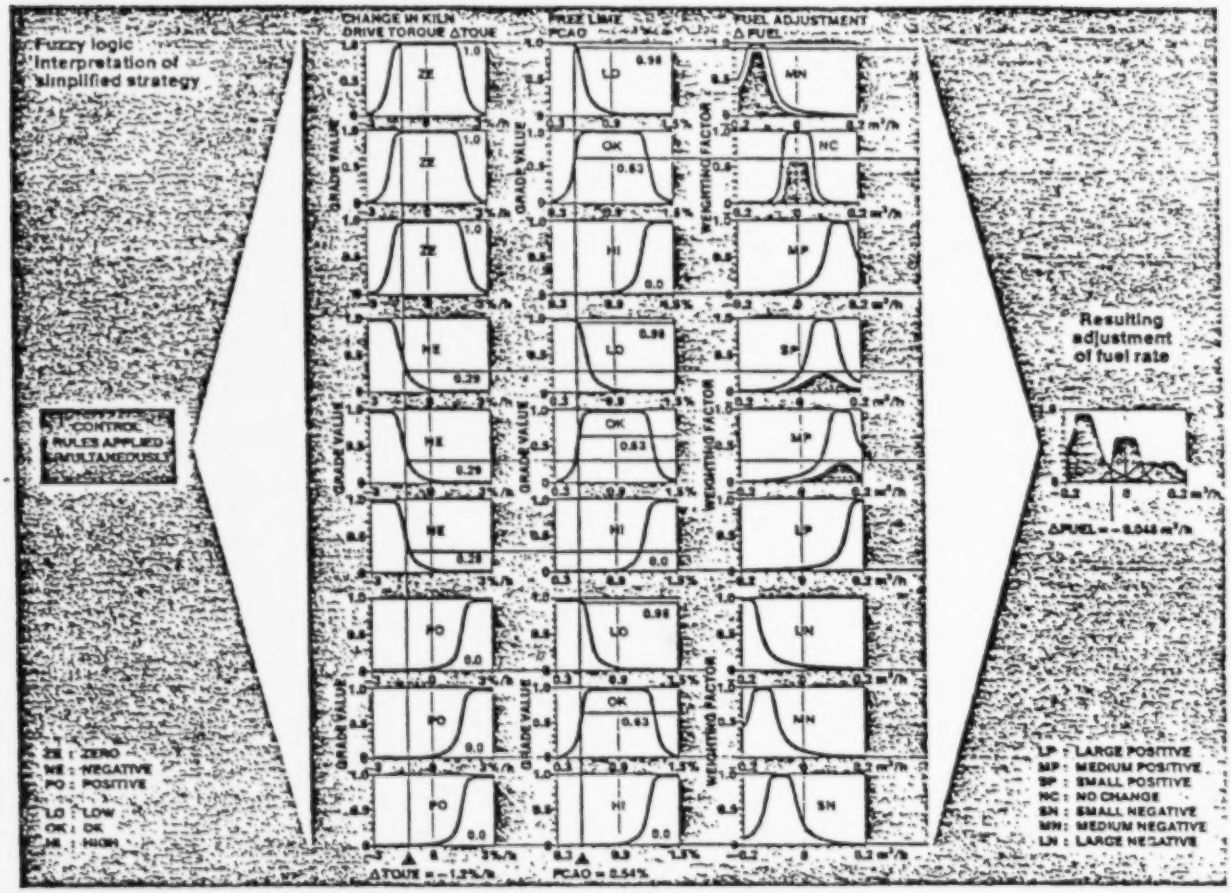


Fig. 4. Graphical illustration of the fuzzy logic interpretation of the nine kiln control rules shown in Fig. 2. The figure illustrates how the rules result in a fuel adjustment equal to -0.048 m³/h when the torque change is -1.2%/h and free lime is equal to 0.54%.

Figure 9 - Fuzzy controller output for a lime kiln.

Fuzzy logic controllers are used today in everything from washing machines to cameras. Their primary advantage is the description of the rules in linguistic terms so that human experts can readily interpret and improve the way that the controller works. Recently, some work has been done on making fuzzy controllers adaptive, so that they change their rules over time based on an optimization criteria that uses measured inputs from the environment [Cox, 1993].

P153167.PDF [Page: 32 of 57]

3.2 Fuzzy Pattern Recognition

Fuzzy pattern recognition uses definitions for class membership that closely resemble fuzzy set membership values. Based on Bezdek's original work [Bezdek, 1981], most fuzzy classifiers use the unsupervised clustering method known as the fuzzy ISODATA algorithm. Using a least squares error minimization technique, the result of fuzzy clustering is that each data point is classified such that it has partial membership in each of the c possible classes. The resulting membership array U has c rows and n columns, where n is the number of data points. For illustration, Bezdek's butterfly data example is shown in Figure 10.

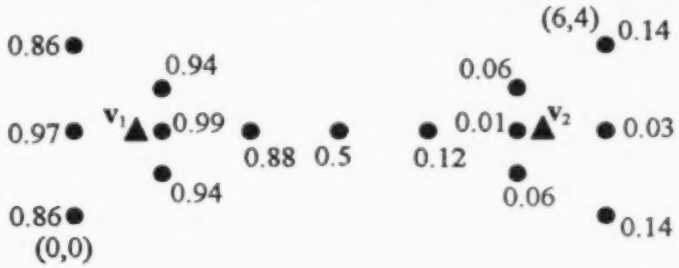


Figure 10 - Fuzzy classification results of the butterfly data example [Bezdek, 1981]

Figure 10 shows the results of classifying 15 points with the fuzzy ISODATA algorithm, using the Euclidean distance metric with c = 2 classes, and the "fuzziness parameter" m = 2. The data is symmetrically distributed over an integer (6,4) space, and the single real numbers next to each data point are the membership values of each data in class 1. The cluster centers $\mathbf{v}_1 = (0.83, 2.00)$ and $\mathbf{v}_2 = (5.14, 2.00)$ are indicated with triangles. The membership array U is given below.

$$U = \begin{bmatrix} .86 & .97 & .86 & .94 & .99 & .94 & .88 & .50 & .12 & .06 & .01 & .06 & .14 & .03 & .14 \\ .14 & .03 & .14 & .06 & .01 & .06 & .12 & .50 & .88 & .94 & .99 & .94 & .86 & .97 & .86 \end{bmatrix}$$

This approach has also been used for image classification, with the result that each classified pixel has a membership value in each class [Di Zenzo et al, 1987, Wang, 1990; Rignot et al, 1992]. Bayesian estimation [e.g. Duda and Hart, 1973] for image classification computes the probability of a given pixel being in each class, and then chooses the class with the largest probability as the pixel class. Of particular importance in determining how well a classification algorithm performed is the "confusion matrix" which shows (for known data) how may pixels were misclassified. As an example, Table 8 is a crop confusion table resulting from a classification of LANDSAT Multi-Spectral Scanner (MSS) data. Classes are labelled C (corn), R (rice), A (alfalfa), S (sugar beets), T (tomatoes) and 0 (background). Numbers in the row labelled 0 indicates pixels from the trainable classes that have been misclassified as belonging to the background.

P153167.PDF [Page: 33 of 57]

Table 8 - Crop confusion table after relaxation for MSS data [from Di Zenzo et al, 1987].

Classified as → Actual	C	R	A	S	T	
0	0	9	0	6	6	21
C	585	84	48	3	0	720
R	3	504	3	0	0	510
A	3	0	432	75	111	621
S	0	3	99	492	36	630
T	9	0	18	24	447	498
	600	600	600	600	600	

Overall classification accuracy: 82.0%

Recent attempts have improved the fuzzy clustering method [e.g. Hathaway and Bezdek, 1994] and have also investigated the incorporation of fuzzy methods into pattern classification using artificial neural networks [e.g. Bezdek, 1992].

WaterMark

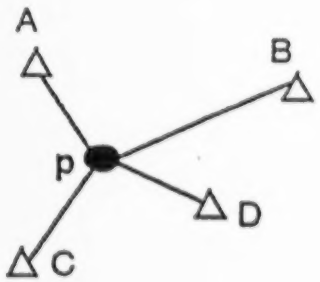
P153167.PDF [Page: 34 of 57]

4. Fuzzy Logic and Geographical Data

Fuzzy logic has been applied to geographical data in several research and development projects.

Lam [1994] describes a method for interpolation of heights based on fuzzy arithmetic. Interpolation of the height for an unknown point combines the heights of its nearest neighbours based on a weighted distance calculation that incorporates an upper and lower uncertainty bound on both the distance to and the height of the known points. The result gives all interpolated points a lower and upper uncertainty bound.

	Elevation (H)	Distance (D)
Point A	10 (9,11)	1 (0.8, 1.2)
Point B	20 (19,21)	2 (1.6, 2.4)
Point C	15 (14,16)	1.5 (1.2,1.8)
Point D	13 (12,14)	1 (0.8, 1.2)



$$H_p = \frac{\Sigma(H/D)}{\Sigma(1/D)}$$

$$H_1 = \frac{10(9,11)/1 + 20(19,21)/2 + 15(14,16)/1.5 + 13(12,14)/1}{1/1 + 1/2 + 1/1.5 + 1/1}$$

$$H_1 = 13.6(12.6, 14.6)$$

$$H_2 = \frac{10(9,11)/1(0.8,1.2) + 20(19,21)/2(1.6,2.4) + 15(14,16)/1.5(1.2,1.8) + 13(12,14)/1(0.8,1.2)}{1/1(0.8,1.2) + 1/2(1.6,2.4) + 1/1.5(1..2,1.8) + 1/1(0.8,1.2)}$$

$$H_2 = 13.6(8.4, 21.9)$$

Figure 11 - Fuzzy nearest-neighbour interpolation example [from Lam, 1994].

Sui [1994] briefly reviews some attempts at incorporating fuzzy logic into geographical information systems. Of note is the prototype soil information retrieval system of Kollias and Voliotis [1991] which allows the user to specify queries about soil parameters using linguistic terms such as "well drained", "quite shallow" or "quite deep". This system was built on top a standard relational database, and transformed the fuzzy query language into standard relational database queries. The results returned from the query are the same as those returned from a standard SQL query with a threshold of 0.6 on the fuzzy set membership grade matching the database contents. Figure 10 shows the membership function for "sufficient potassium" and Figure 11 shows a sample query with the resultant found soil profile. A user-defined threshold of 0.8 has been used in the example in Figure 11.

P153167.PDF [Page: 35 of 57]

Sufficient potassium = (1/200 + 0·96/190 + 0·92/180 + 0·90/170 + 0·86/160
+ 0·8/150 + 0·73/140 + 0·67/130 + 0·60/120 + 0·55/110
+ 0·5/100 + 0·45/90 + 0·40/80 + 0·35/70 + 0·28/60
+ 0·20/50 + 0·16/40 + 0·10/30)

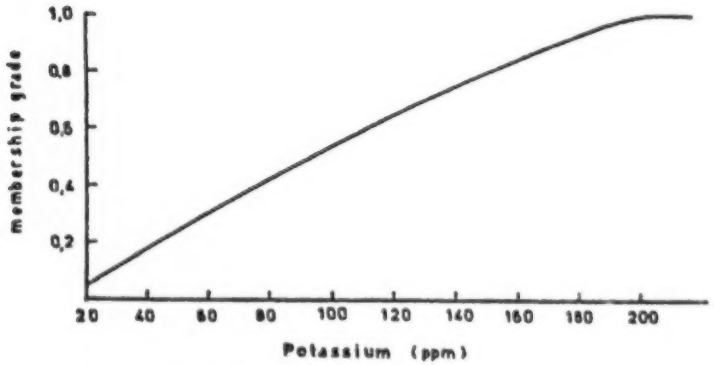


Figure 12 - The membership function for fuzzy set "sufficient potassium" [from Kollias and Voliotis, 1991].

range of *p* is PROFILE
range of *c* is CHEMICAL
fretrieve into QU4 (*p*.all) where *p*.PNO = *c*.PNO
and *p*.HORNO = *c*.ORD
and *p*.ALTITUDE GREATER THAN 800
and *p* is WELL_DRAINED
and *c* is QUITE DEEP
with possibility 0·8

PNO	HORNO	ORDER	SLOPE	DRAINAGE	ALTITUDE
P5	5	ALFISOL	2	B	890

Figure 13 - Fuzzy query and result [from Kollias and Voliotis, 1991].

The linguistic modifier "quite" for a fuzzy set A gives a new fuzzy set "quite A" defined as

$$\mu_{\text{quite A}}(x) = (\mu_A(x))^{1.5} \qquad (14)$$

P153167.PDF [Page: 36 of 57]

One particular difficulty with most fuzzy logic systems is determining the precise meaning of the linguistic terms used to represent data. Fisher and Orf [1991] performed an experimental study on the campus of Kent State University to determine the meaning of the linguistic terms "near" and "close" (for distances to other buildings from the student centre) among the students they interviewed. A total of 83 students took part, and the results indicate that there were three different clusters of results as to the meaning of "near" and "close". This is illustrated in Figure 14.

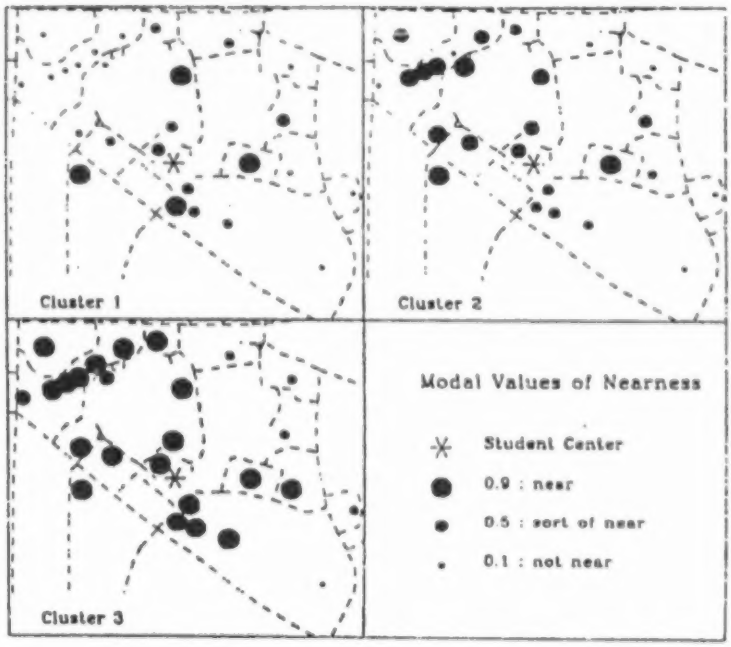
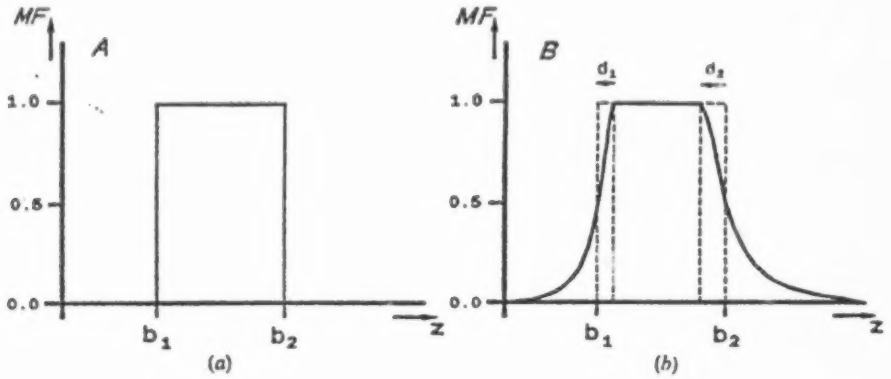


Figure 14 - Maps of the modal fuzzy membership values for each of the three clusters identified for the concept "near" to the Student Center [from Fisher and Orf 1991].

No statistically significant indicator was found which could predict which cluster a student would belong to. The results show that it would be very difficult to enable a geographical query of the type "What buildings are **near** the student centre?" that would satisfy all of the student population. Some individuals would accept almost any building on campus, whereas others would accept only a restricted list. If one was to build a system using such fuzzy query definitions, it seems that the user should have control over the meaning of linguistic "fuzzy" terms so that the query results are meaningful.

Heuvelink and Burrough [1993] describe a system for error propagation using Boolean logic and continuous classification. Their continuous classification uses the fuzzy set notion of partial membership. Figure 15 shows the Boolean and continuous membership functions used in the paper.

P153167.PDF [Page: 37 of 57]



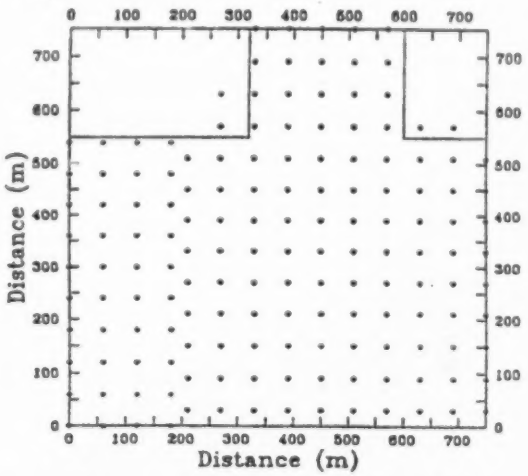
$$MF^C(z) = \frac{1}{1+\left(\frac{z-b_1-d_1}{d_1}\right)^2} \quad \text{if } z < b_1 + d_1$$

$$MF^C(z) = 1 \quad \text{if } b_1 + d_1 \leqslant z \leqslant b_2 - d_2$$

$$MF^C(z) = \frac{1}{1+\left(\frac{z-b_2+d_2}{d_2}\right)^2} \quad \text{if } z > b_2 - d_2$$

Figure 15 - Boolean (a) and continuous (b) membership functions for class membership [from Heuvelink and Burrough 1993].

To illustrate their method, they used soil sample profiles (154 in total) in an approximately 55 hectare area (about 750 m by 750 m) in Alberta. The data measured from the profiles included percent sand (SAND), percent clay (CLAY) and the natural logarithm of soluble sodium content of the C horizon (70-80 cm. deep) (LNNA). The results of the original soil survey are shown in Figure 16.



Lacombe Extension area

Table 1. Summary statistics of the soil data.

	Mean	sd	CV	Min	Max	Correlation with LNNA	Correlation with SAND	Correlation with CLAY
LNNA	2·14	1·15	53·9	0·26	4·38	1·00	−0·30	0·36
SAND	21·9	20·0	91·5	0·0	65·0	−0·30	1·00	−0·85
CLAY	41·1	9·3	22·5	21·0	72·0	0·36	−0·85	1·00

Figure 16 - Original soil data from Alberta [from Heuvelink and Burrough 1993].

P153167.PDF [Page: 38 of 57]

This data was interpolated on a 5 m by 5 m grid using the Kriging interpolation method to arrive at the results shown in Figure 17.

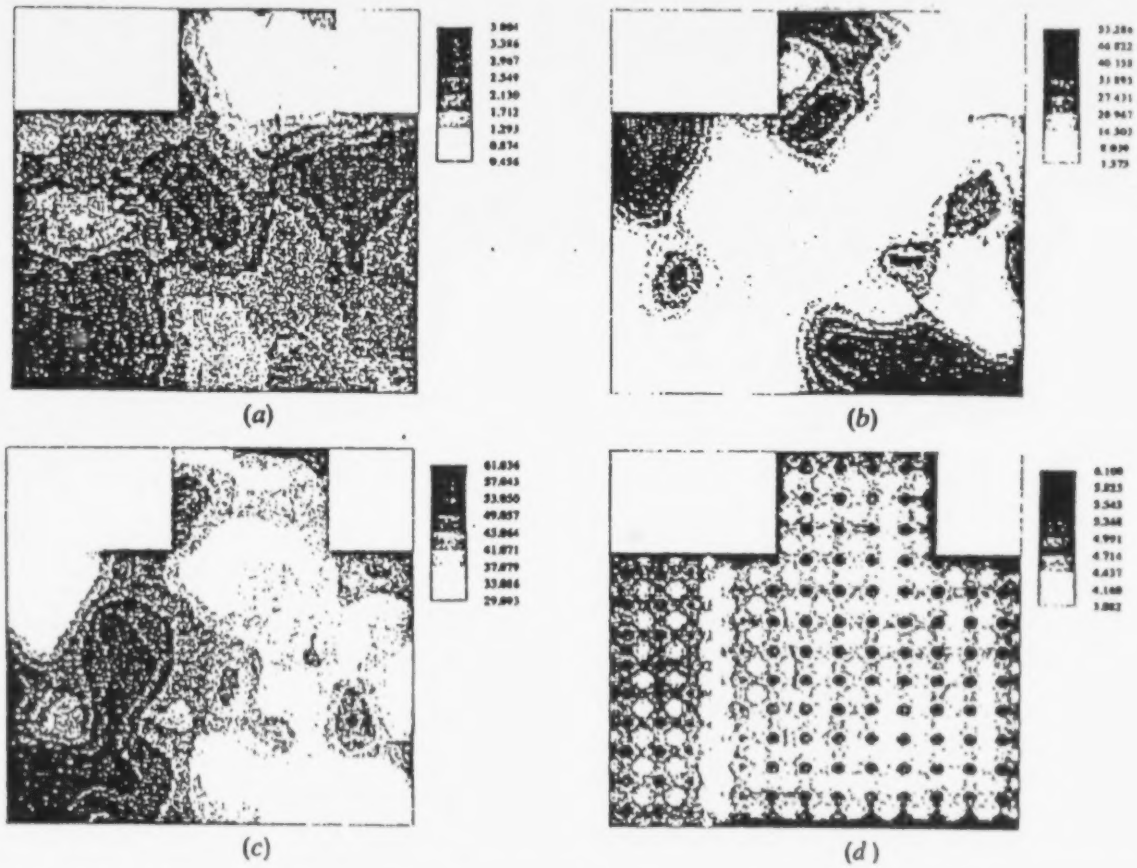


Figure 4. Maps of the three soil attributes produced by ordinary block co-kriging to a 5 m by 5 m cell size: (*a*) LNNA (ln (meq 100 g^{-1})), (*b*) SAND (per cent), (*c*) CLAY (per cent), (*d*) co-kriging block standard error for CLAY (per cent). (Note all error maps have the same form but different ranges.)

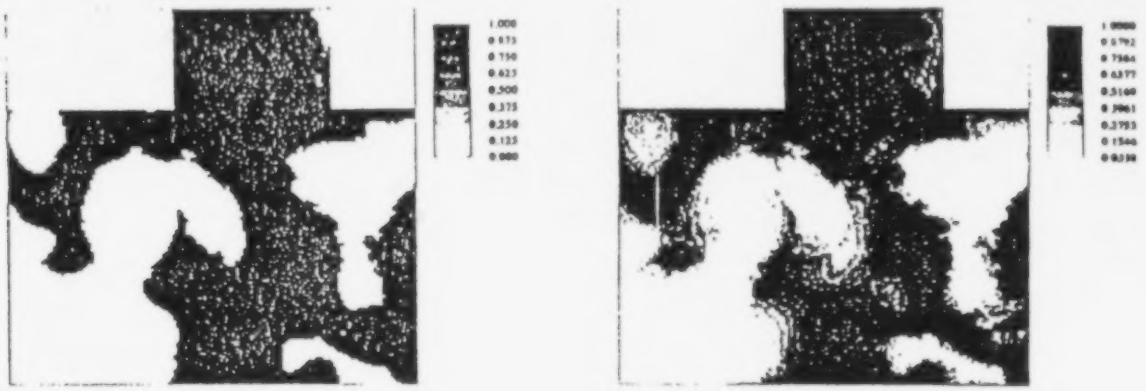
Figure 17 - Interpolated soil data [from Heuvelink and Burrough 1993].

With a three raster datasets in hand, one can now perform queries such as the following:

$$10\% \leq \text{SAND} \leq 15\% \textbf{ AND } 15\% \leq \text{CLAY} \leq 50\% \textbf{ AND } \text{LNNA} \leq 2.5 \qquad (15)$$

P153167.PDF [Page: 39 of 57]

This corresponds approximately to those soils suitable for agriculture. The AND operation is performed using the Zadeh definition of minimum. The results are shown in Figure 18, and shows the partial memberships resulting from the continuous classification method.



(*a*) (*b*)

Figure 18 - Boolean (a) and continuous (b) classification of the soils according to the suitability criteria in equation (15) [from Heuvelink and Burrough 1993].

4.1 Extension Principle

The extension principle states [e.g. Klir and Folger, 1988; Zadeh, 1975] that given any function f mapping points in set X to points in set Y, and any fuzzy set A, where

$$A = \mu_1/x_1 + \mu_2/x_2 + \ldots + \mu_n/x_n$$

then

$$f(A) = f(\mu_1/x_1 + \mu_2/x_2 + \ldots + \mu_n/x_n) = \mu_1/f(x_1) + \mu_2/f(x_2) + \ldots + \mu_n/f(x_n) \quad (16)$$

Thus, any function defined for crisp sets can be extended to fuzzy sets. As an example, consider the fuzzy set A = small, defined as

$$A = \text{small} = 1/1 + 1/2 + 0.8/3 + 0.6/4 + 0.4/5 + 0/6 + \ldots + 0/10 \quad (17)$$

Then, we can write

$$(\text{small})^2 = 1/1 + 1/4 + 0.8/9 + 0.6/16 + 0.4/25 + 0/36 + \ldots + 0/100 \quad (18)$$

P153167.PDF [Page: 40 of 57]

4.2 Extensions to Two Dimensions

One can generalize notions about fuzzy sets to fuzzy points, which can then be extended to the notion of a fuzzy line. A fuzzy point is defined as [Xu and Nickerson, 1995] a convex normalized fuzzy set A of the plane C such that

(a) there exists exactly one $c_0 \in C$ satisfying $\mu_A(c_0) = 1$, and
(b) μ_A is piecewise continuous.

Figure 19 is an illustration of two fuzzy points at positions (2, 6) and (8, 7).

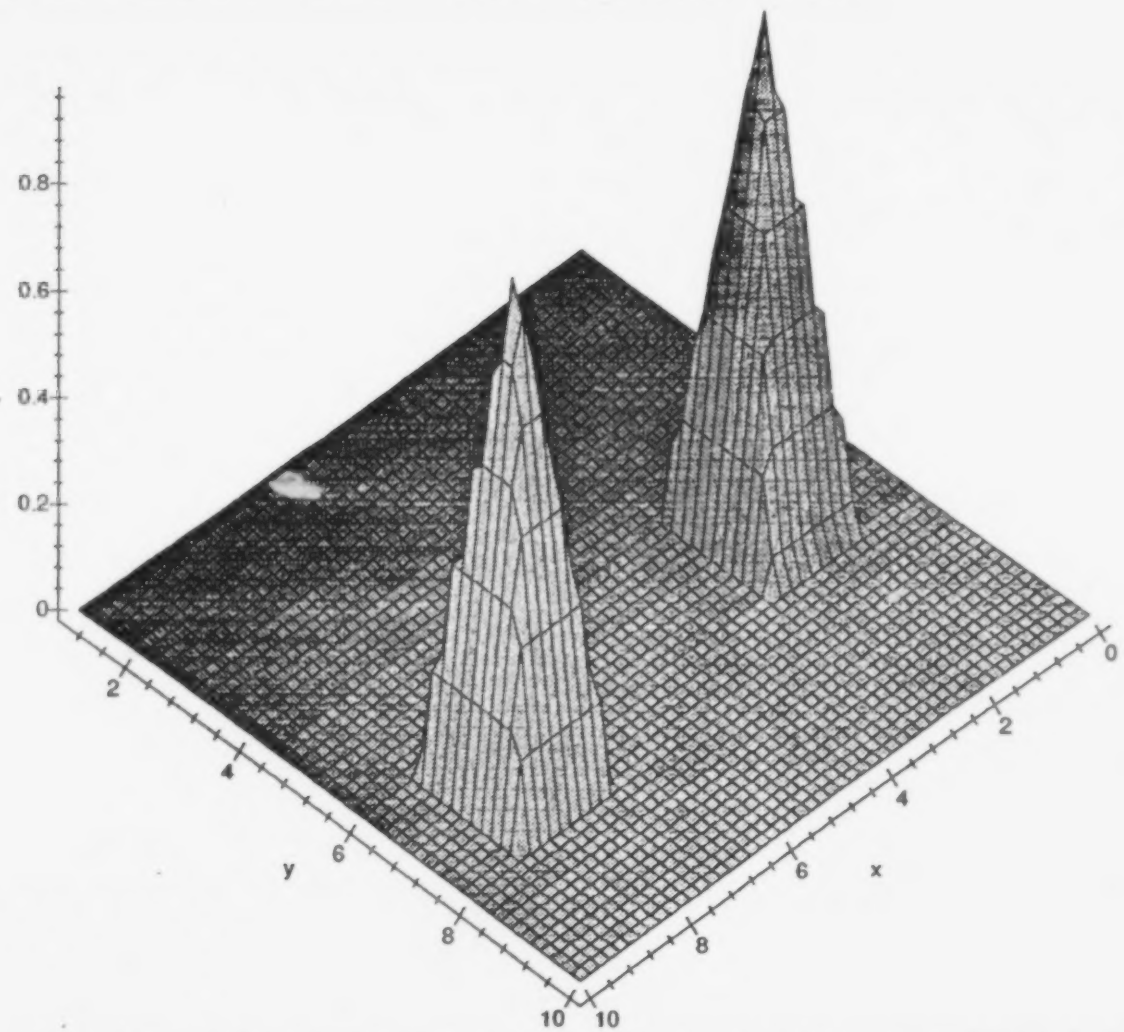


Figure 19 - Plot of two fuzzy points at locations (2, 6) and (8, 7) [from Xu and Nickerson, 1995].

WaterMark

P153167.PDF [Page: 41 of 57]

A fuzzy line segment between two fuzzy points A and B can be expressed as the set of fuzzy points [Xu and Nickerson, 1995]

$$[A, B] = \{\alpha A + (1 - \alpha)B : \alpha \in [0,1]\}, \quad (19)$$

where

$$\mu_{\alpha A + (1-\alpha)B} = \alpha\, \mu_A + (1 - \alpha)\, \mu_B \quad (20)$$

A collection of n fuzzy line segments forms the notion of a fuzzy polygon, from which we can determine a fuzzy area. One could base uncertainty viewing methods on these notions of fuzzy points and fuzzy line segments.

A fuzzy circle O(A,r), with fuzzy point A at the centre and radius r is defined as

$$\mu_{O(A,r)}(C) = \sup_{d(u,v)=r} \min(\mu_A(u), \mu_C(v)) \quad (21)$$

where C is a fuzzy point. O(A,r) is a fuzzy set on C.

Where rules about uncertainty management can be established, fuzzy logic could be used to make inferences about uncertainty in geographical information based on these rules. For example, one rule might be

IF *most* attribute level data has a *good* source quality, and
most feature level data has a *good* source quality,
THEN dataset source quality is *good*

Fuzzy logic can be used to assist in the propagation of uncertainty using rules of this type.

WaterMark

P153167.PDF [Page: 42 of 57]

5. Suggested Modifications to DIGEST for Uncertainty Management

We have conducted a market and literature survey in Phase 1 of this project, which revealed an interest on uncertainty management but a lack of uncertainty information and support from GIS vendors. We have also surveyed the use of DIGEST and its capability to handle uncertainty. The results of this survey lead to the recommendations given below.

In this report, we have surveyed the state of technology in uncertainty data collection, modelling, representation, propagation, visualization, and the application of fuzzy logic in handling spatial data. It is clear from the survey that these are active areas of research, and some of the findings can be implemented in systems to manage uncertainty of spatial data.

Another objective of this study is to recommend changes to DIGEST in order to facilitate uncertainty management. In making the recommendations, we have regarded DIGEST as both an interchange standard and a data browsing system, with the possibility of being extended to become a GIS. Indeed, a variation of DIGEST has already been used in the data browsing system of DCW.

When DIGEST is used as an interchange standard, we are concerned with its flexibility of supporting other data models and the ease of writing data conversion software between DIGEST and other formats. When DIGEST is used as a data browsing system, we are concerned with its extensibility to suit particular applications and the ease of writing data management, processing, analysis, and visualization modules.

The following recommendations emphasize resolving some of the internal inconsistencies of DIGEST, improving its flexibility, and enhancing its capability to handle data uncertainty. The inconsistencies of DIGEST are mainly caused by the mixture of two somewhat different designs. This is addressed in Recommendation 5.1. The lack of a modelling language in DIGEST will affect the use of it in both data interchange and data modelling in a GIS environment. This is addressed in Recommendation 5.2. DIGEST provides about 30% of all data quality descriptors we have collected from various sources. We have made several recommendations (5.3, 5.4, 5.5, and 5.6) to extend it to address the findings of our survey in Phase 1. A concern from the users as revealed in our Phase 1 study is the lack of data on uncertainty. Recommendation 5.7 addresses that problem.

5.1 Resolve the Internal Consistencies of DIGEST

DIGEST is composed of a data model, a logical data schema, and a number of encapsulation methods, including ISO 8211, ISO 8824/25 and VRF (Figure 20). The three encapsulation methods are described in Annex A, Annex B, and Annex C of DIGEST Part 2 respectively. The three methods have thus been known as DIGEST A, DIGEST B, and DIGEST C.

P153167.PDF [Page: 43 of 57]

Product specifications are DIGEST profiles. They assemble the appropriate elements from the DIGEST data model, logical schema, encapsulation rules, and other necessary information such as scale to define the specification of a particular product.

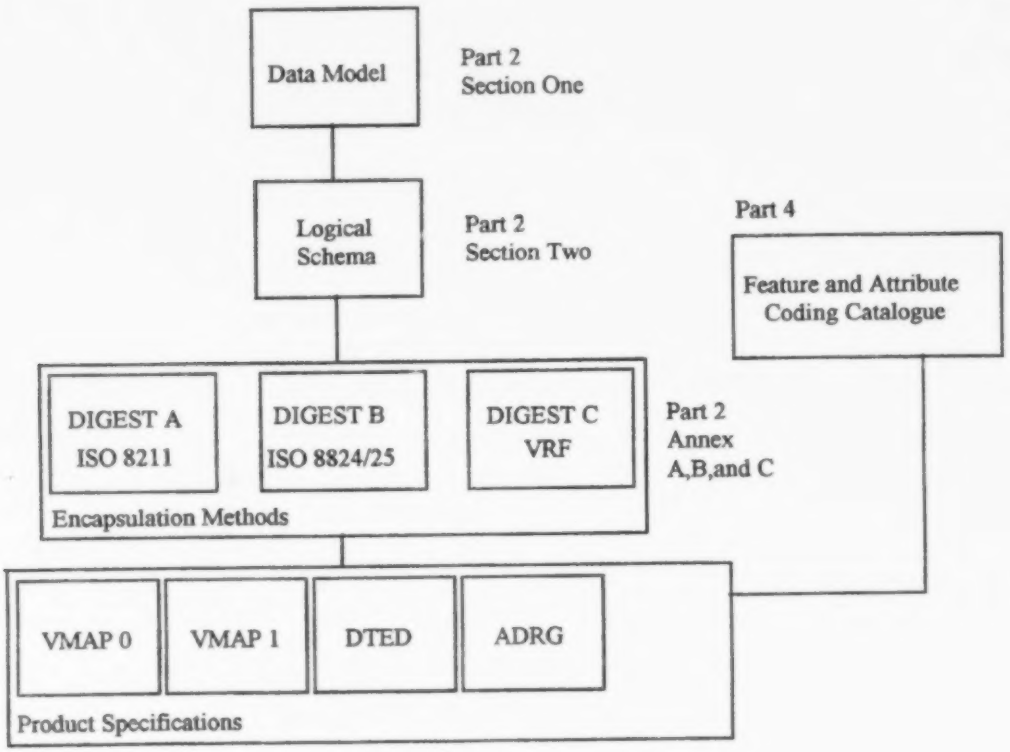


Figure 20 - DIGEST

DIGEST A and DIGEST B are based directly on the logical schema of DIGEST. However, DIGEST C deviates somewhat from the data model and the logical schema for historical reasons. For example, DIGEST C does not recognize raster and matrix data types but DIGEST A and B do.

In handling data quality, DIGEST A and B conform to the Quality File in the Geo Data Subset set out in Part 2 of the DIGEST specification. These descriptors in the Quality File can be associated with a volume, a dataset, a feature, a topological primitive, or an attribute.

DIGEST C places the data quality descriptors in Data Quality Tables that can be stored at the coverage, library, or database levels. A coverage is a collection of topologically related feature classes. A library is used to organize coverages, spatially registered and be in the same coordinate system, in a geographic region. A database is a collection of libraries. Non-standard tables and annotation text can also be created to carry data quality descriptors.

P153167.PDF [Page: 44 of 57]

There are some slight differences between the data qualifiers in DIGEST A/B and C as shown in the following table.

A,B,C	Feature completeness percent
A,B,C	Attribute completeness percent
A,B,C	Logical consistency
A,B,C	Edition number of data set
A,B,C	Date of creation of data set
A,B,C	Date of revision of data set
A,B	Recompilation count
A,B	Revision count
A,B,C	Name of product specification (e.g., DCW)
A,B,C	Date of product specification
A,B	Specification amendment
A,B,C	Date of earliest source
A,B,C	Date of latest source
A,B,C	Standard deviation of quantitative attributes
A,B,C	Percent reliability of qualitative attributes
A,B	Name of collection specification
C	Name of included source file
A,B,C	Absolute horizontal accuracy
A,B,C	Unit of measure for absolute horizontal accuracy
A,B,C	Absolute vertical accuracy
A,B,C	Unit of measure for absolute vertical accuracy
A,B,C	Relative horizontal accuracy
A,B,C	Unit of measure for relative horizontal accuracy
A,B,C	Relative vertical accuracy
A,B,C	Unit of measure for relative vertical accuracy
A,B,C	Miscellaneous comments - free text for non-standard information

In DIGEST A and B, the positional accuracy descriptors can be associated with a region delineated by a polygon. DIGEST C recognizes data quality coverages, which are special coverages of level 3 topology (full topology) containing faces and edges to help describe data quality. For instance, a face can be created in the data quality coverage to delineate an area with a uniform data quality description such as a part of the database with data came from a single source. An edge can be used in the data quality coverage to show where two maps meet, and a data quality descriptor can hence be associated with the edge to show that features on both sides of the line do not all meet.

For DIGEST C, lineage information is carried in a text narrative table called LINEAGE.DOC. The contents of the lineage table is not standardized although it is recommended in Section C5.2.1 that at a minimum, it should contain information on processing tolerances, interpretation rules applied to source materials, and basic production and quality assurance procedures/

WaterMark

P153167.PDF [Page: 45 of 57]

DIGEST C appears to be a competing standard within DIGEST in that it has its own data model. For instance, it recognizes tiles, winged edges, and spatial indices that are not part of the theoretical DIGEST data model as described in Part 2, Section One of the DIGEST specification. To further add to the confusion, many of the product specifications, such as VMAP0 and VMAP1, are based on DIGEST C or its variant VPF.

WaterMark Recommendation: Harmonize the Various DIGEST Encapsulations

In the previous section, we have described the inconsistencies between DIGEST C and the logical schema of DIGEST. They are not major but are sources of confusion. It is recommended that the differences be resolved as soon as possible.

5.2 Modelling Language Extension

WaterMark Recommendation: Develop a Modelling Language for DIGEST

A modelling language for DIGEST can help the users model data to describe their database organizations (if DIGEST were to be used as a framework for GIS) and to help developers write translators between DIGEST and other systems (if DIGEST were to be used as an interchange standard). A properly designed modelling language can highlight the similarities and differences between the various data types, make the documentations even more structured, concise, and readable. The use of the extended Structured Systems Analysis and Design (SSADM) Entity-Relationship Modelling method as shown in Part 2 Annex E is a good start. It is recommended that a modelling language for DIGEST be developed.

5.3 Specifications for Quality Control Testing

Standards must allow for the complete specification of quality control testing performed on a particular dataset. For example, Bartakovitch [1994] outlines the quality controls required for paper map products of the Canada Centre for Mapping. Parameters evaluated include integrity, accuracy, lineage, and process precision. An acceptable quality level is established through stratified random sampling and statistical sampling of produced maps. A determination is made of the number of critical, major and minor defects in the map. Map sheets with ≥ 1, 2 or 20 of these defect types, respectively, cause them to be unacceptable.

At least 70 points are sampled for each map sheet to determine the planimetric accuracy of a map sheet. At least 140 points are sampled for each map sheet to determine overall relief accuracy for a map sheet. Complete specifications for Canadian NTS maps are given in EMR [1994].

P153167.PDF [Page: 46 of 57]

SAIF, SDTS, and the FGDC metadata standard all include specifications for quality control testing. As shown in the complete list of data qualifiers in a previous section, test reports on method and result are included for accuracy, logical consistency, completeness, and lineage. DIGEST do not follow the same practice except for the logical consistency description. The lineage table in DIGEST C could be used for this purpose, but it is an unstructured narrative table. That means information about quality control testing are not explicit and are not standardized.

WaterMark Recommendation: Add Specifications for Quality Control Testing

It is recommended that the DIGEST standard be able to record the quality assurance test results carried out on a specific DIGEST dataset. These test results will probably be similar in nature to the results of the EMR quality assurance testing for paper map sheets.

5.4 Data Processing Errors

Data processing errors arise from many sources. The lineage data of the DIGEST data quality representation should contain explicitly a complete representation of the processing carried out to arrive at the data in a dataset.

WaterMark Recommendation: Modify the lineage data quality representation in DIGEST

Data processing errors arise from many sources. The lineage data of the DIGEST data quality representation should contain a complete representation of the processing carried out to arrive at the data in a dataset

5.5 Enhance Topological Consistency Description

DIGEST currently supports a general statement on logical consistency. It is recommended that the topological consistency part of it be specifically identified because topological inconsistencies can sometime have more impact on data quality than positional inaccuracies

WaterMark Recommendation: Enhance Topological Consistency Description

Topological inconsistencies can be detected by tracing specified topological links (in data that has level 1, 2 or 3 topology) to determine if there are missing links, or if the links are inconsistent. Provision for placing quality assurance measures of this type in the data quality section should be made. A specification of what type of test was made and the results should be given.

P153167.PDF [Page: 47 of 57]

5.6 Uncertainty for Raster Data

Based on the discussions about raster data. It is recommended that DIGEST be able to carry a representation of the probability value or membership value associated with the class it has been assigned. This is for raster data resulting from an image classification algorithm.

WaterMark Recommendation: DIGEST carry a probability value

Based on the discussions about raster data above, it is recommended that DIGEST be able to carry a representation of the probability value or membership value associated with the class it has been assigned. This is for raster data resulting from an image classification algorithm.

5.7 Toolkits

The availability of useful toolkits can enhance a standard, and make it easier for users to understand and accept.

WaterMark Recommendation: Toolkits for generating quality assurance measures be developed

It is recommended that toolkits for generating quality assurance measures from existing DIGEST datasets be developed.

WaterMark Recommendation: Toolkits for viewing uncertainty information be developed

Without effective means of viewing known uncertainty information, users are often unaware of the uncertainty. Simple tools for viewing uncertainty information should be developed.

P153167.PDF [Page: 48 of 57]

6. Conclusions

The suggestions for modifications to DIGEST mentioned above are cross-referenced with the issues outlined in Chapter 1 in the following Table.

Table 9 - Cross Reference of Suggested DIGEST Modifications

Suggested Modification	DIGEST Issue(s) from Table 1	User Issue(s) from Table 2
5.1 Different DIGEST encapsulations	I	
5.2 Complete specification of quality control testing	II, III	1, 4
5.3 Uncertainty for raster data	V	6, 7
5.4 Toolkits	VII	3, 8, 9
5.5 Data processing errors	IV	2, 5
5.6 Uncertainty in topological consistency	VI	
5.7 User extendible model		4

Since there is a need to freeze DIGEST, changes can not be easily made to it for the time being. That could mean that recommendations 5.1, 5.3, 5.4, 5.5, and 5.6 cannot be implemented in the near future. Recommendation 5.2 does not require changes to DIGEST and could lead to a wider acceptance of it. Recommendation 5.7 also does not require changes to DIGEST. Making more data in the DIGEST format containing descriptors of data quality and provide software to view such information will show the user community that such information can indeed be obtained and that their incorporation into GISs can indeed be implemented. We therefore suggest that recommendations 5.2 and 5.7 be considered immediately and plans be made to implement the other recommendations as soon as possible.

With these modifications, DIGEST will be more *"user friendly"* and will be improved in its capability to clearly and precisely specify uncertainty in the geographical data it represents. For DIGEST to become more widely used and accepted, it must overcome the marketplace perception that it is *"closed"* and difficult to use. Managing the data uncertainty will be greatly assisted by the construction of uncertainty management toolkits.

P153167.PDF [Page: 49 of 57]

7. References

Altman, D. "Fuzzy set theoretic approaches for handling imprecision in spatial analysis", *Int. J. Geographical Information Systems*, vol.8, no.3, 1994, pp.271-289.

Amrhein, C.G. and Griffith, D.A. "A model for statistical quality control of spatial data in a GIS", Proc. of the Canadian Conference on GIS, Ottawa, March 18-21, 1991, pp.91-103.

Antenucci, J.C., K. Brown, P.L. Croswell, M.J. Kevany and H. Archer (1991) *Geographic Information Systems: A guide to the technology*. Van Nostrand Reinhold, New York.

Bartakovich, A. "Quality assurance of National Topographic System paper products", paper, unknown source, 1994(?)

Banai, R. "Fuzziness in geographical information systems: contributions from the analytic hierarchy process", *Int. J. Geographical Information Systems*, vol.7, no.4, 1993, pp.315-329.

Bezdek, J.C. "Computing with uncertainty", *IEEE Communications Magazine*, Sept. 1992, pp.24-36.

Bezdek, J.C. *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, NY, 1981.

Bolstand, P.V., Gessler, P. and Lillesand, T.M. "Positional uncertainty in manually digitized map data", *Int. J. Geographical Information Systems*, vol.4, no.4, 1990, pp.399-412.

British Columbia, "Spatial Archive and Interchange Format: Formal Definition", Release 3.2, Surveys and Resource Mapping Branch, Ministry of Environment, Lands and Parks, Province of British Columbia, Canada; also Canadian General Standards Board publication CAN/CGSB-171.1-95; available from anonymous ftp site ftp.crl.gov.bc.ca, file pub/saif3/saif32ps.zip, Jan. 1995.

Canadian General Standards Board, "Technical Criteria for the Evaluation of Existing Interchange Standards", Committee on Geomatics, Ottawa, Mar. 16, 1991.

Cox, E. "Adaptive fuzzy systems", *IEEE Spectrum*, Feb. 1993, pp.27-31.

Cox, E. *The Fuzzy Systems Handbook*, AP Professional, Boston, 1994.

Defense Mapping Agency, *Digitizing the Future*, Fourth edition, Fairfax, VA, 1994(?).

Department of National Defence, "The Digital Geographic Information Exchange Standard (DIGEST) Parts 1, 2 and 3", Directorate of Geographic Operations, Canada, Edition 1.2, January 1994.

WaterMark

P153167.PDF [Page: 50 of 57]

Department of National Defence, "The Digital Nautical Chart Draft Military Specification", MIL-D-89023, Directorate of Geographic Operations, Canada, DRAFT Oct. 19, 1993.

Derenyi, E., Y.C. Lee, G.W. Moore, and A. Chow (1993). "The speed and accuracy of semi-automated digitizing." *Proceedings of the 7th Annual Symposium on Geographic Information Systems in Forestry, Environment and Natural Resources Management, Vol. 2*, February 15-18, Vancouver, B.C., pp. 1011-1019,

Di Zenzo, S., Bernstein, R., Degloria, S.D. and Kolsky, H.G. "Gaussian maximum likelihood and contextual algorithms for multicrop classification", *IEEE Trans. on Geoscience and Remote Sensing*, vol.GE-25, no.6, Nov. 1987, pp.805-814.

Dubois, D. and Prade, H. *Fuzzy Sets and Systems*, Academic Press, NY, 1980.

Duda, R.O. and Hart, P.E. *Pattern Classification and Scene Analysis*, John Wiley and Sons, NY, 1973.

Dunn, R., A.R. Harrison, and J.C. White (1990) "Positional accuracy and measurement error in digital databases of land use: An empirical study." *International Journal of Geographical Information Systems*, Vol.4, No.4, pp. 385-398.

Dutton G. (1989) "Modeling locational uncertainty via hierarchical tessellation." In: *Accuracy of Spatial Databases*, Eds. Goodchild, M.F. and S. Gopal, Taylor and Francis, London. pp. 125-140.

Edwards, G. "Error minimization in integrated GIS/IAS system design", Proc. of the 5th Int. Symposium on Spatial Data Handling, vol.1, Aug. 3-7, 1992, pp.20-29.

Egenhofer, M.J. and Sharma, J. "Topological consistency", Proc. of the 5th Int. Symposium on Spatial Data Handling, Charleston, South Carolina, vol.2, Aug. 3-7, 1992, pp.335-343.

Elkan, C. "The paradoxical success of fuzzy logic", *IEEE Expert*, August 1994, pp.3-8.

Energy, Mines and Resources Canada, "Accuracy", Canada Centre for Mapping, Surveys Mapping and Remote Sensing Sector, DRAFT standard, 1994(?)

Fisher, P. "Real-time randomization for the visualization of uncertain spatial information", Proc. of the 5th Int. Symposium on Spatial Data Handling, Charleston, South Carolina, vol.2, Aug. 3-7, 1992, pp.491-494.

Fisher, P.F. and Orf, T.M. "An investigation of the meaning of near and close on a university campus", *Comput., Environ. and Urban Systems*, Vol.15, 1991, pp.23-35.

Goodchild, M.F. "Geographical data modeling", *Computers & Geosciences*, vol.18, no.4, 1992, pp.401-408.

WaterMark

P153167.PDF [Page: 51 of 57]

Goodchild, M.F. and Min-hua, W. "Modeling error in raster-based spatial data", Proc. of the 3rd Int. Symposium on Spatial Data Handling, Sydney, Australia, Aug. 17-19, 1988, pp.97-106.

Hathaway, R.J. and Bezdek, J.C. "NERF c-means: non-euclidean relational fuzzy clustering", *Pattern Recognition*, vol.27, no.3, 1994, pp.429-437.

Heuvelink, G.B. and Burrough, P.A. "Error propagation in cartographic modelling using Boolean logic and continuous classification", *Int. J. Geographical Information Systems*, vol.7, no.3, 1993, pp.231-246.

Heuvelink, G.B., Burrough, P.A. and Stein, A. "Propagation of errors in spatial modelling with GIS", *Int. J. Geographical Information Systems*, vol.3, no.4, 1989, pp.303-322.

Hunter, G.J. "Techniques for assessing the effect of processing errors in spatial databases", Proc. of the Annual Conference of Urban and Regional Information Systems Association (URISA), vol.1, July 12-16, 1992, pp.24-33.

Hunter, G.J. and M.F. Goodchild (1993). "Managing uncertainty in spatial databases: putting theory into practice." *1993 URISA Proceedings*, Vol.1, pp. 1-14.

Klir, , G.J. and Folger, T.A. *Fuzzy Sets, Uncertainty and Information*, Prentice Hall, Englewood Cliffs, NJ, 1988.

Kollias, V.J. and Voliotis, A. "Fuzzy reasoning in the development of geographical information systems", *Int. J. Geographical Information Systems*, vol.5, no.2, 1991, pp.209-223.

Krishnapuram, R., Keller, J.M. and Ma, Y. "Quantitative analysis of properties and spatial relations of fuzzy image regions", *IEEE Trans. on Fuzzy Systems*, vol.1, no.3, 1993, pp.222-233.

Lagos, R., Avendano, A., Prado, V., Horwitz, I., Wasserman, S., Losonsky, G., Cryz, S., Kaper, J., and Levine, M. "Attenuated live choleral vaccine strain CVD 103-HgR elicits significantly higher serum vibriocidal antibody titers in persons of blood group O", *Infection and Immunity*, vol.63, no.2, Feb. 1995, pp.707-709.

Lam, S. "Fuzzy sets advance spatial decision analysis", *GIS World*, Dec. 1993, pp.58-59.

Lam, S. "Uncertainty propagation in surface interpolation with fuzzy arithmetic", Proc. of the GIS'94 Symposium, Vancouver, B.C., Feb. 1994, pp.367-372.

Lanter, D.P. and Veregin, H. "A research paradigm for propagating error in layer-based GIS", *Photogrammetric Engineering & Remote Sensing*, vol.58, no.6, June, 1992, pp.825-833.

WaterMark

P153167.PDF [Page: 52 of 57]

Lee, C.C. "Fuzzy logic in control systems - parts 1 and 2", *IEEE Transactions on Systems, Man and Cybernetics*, vol.20, no.2, Mar. 1990, pp.404-435.

Lee, T., Richards, J.A. and Swain, P.H. "Probabilistic and evidential approaches for multisource data analysis", *IEEE Trans. on Geoscience and Remote Sensing*, vol.GE-25, no.3, May 1987, pp.283-293.

Leung, Y., Goodchild, M.F. and Lin, C.C. "Visualization of fuzzy scenes and probability fields", Proc. of the 5th Int. Symposium on Spatial Data Handling, Charleston, South Carolina, vol.2, Aug. 3-7, 1992, pp.480-490.

Martin-Clouaire, R. "Semantics and computation of the generalized modus ponens: the long paper", *Int. J. of Approximate Reasoning*, vol.3, 1989, pp.195-217.

McKellar, D.G. "The global geospatial information and services initiative", GGIS Vision Paper Draft, Jan. 25, 1995.

Mikhail, E.M. *Observations and Least Squares*, IEP - A Dun-Donnelley Publisher, NY, 1976.

NATO, "Standardization Agreement (STANAG), Evaluation of Land Maps, Aeronautical Charts and Digital Topographic Data", North Atlantic Treaty Organization, 5th edition, 1989.

Papoulis, A. *Probability, Random Variables, and Stochastic Processes*, 3rd edition, McGraw-Hill, Inc., NY, 1991.

Parratt, L G. (1971) *Probability and Experimental Errors*. Dover Publications, Inc, New York.

Rignot, E., Chellappa, R. and Dubois, P. "Unsupervised segmentation of polarimetric SAR data using the covariance matrix", *IEEE Trans. on Geoscience and Remote Sensing*, vol.30, no.4, July 1992, pp.697-705.

Slama, C.C. (Ed.) (1980). *Manual of Photogrammetry*, Fourth Edition. America Society of Photogrammetry.

Smith, B. "Six-sigma design", *IEEE Spectrum*, vol.30, no.9, Sept. 1993, pp.43-47.

Sugeno, M. (editor) *Industrial Applications of Fuzzy Control*, North-Holland, Amsterdam, 1985.

Sui, D.Z. "Fuzzy logic can help GIS cope with reality", *GIS World*, vol.7, no.9, Sept. 1994, pp.50-53.

Thapa, K. and Bossler, J. "Accuracy of spatial data used in geographic information systems", *Photogrammetric Engineering & Remote Sensing*, vol.58, no.6, June, 1992, pp.835-841.

USGS, Spatial Data Transfer Standard, available from WWW site http://nsdi.usgs.gov/nsdi, 1994a.

WaterMark

P153167.PDF [Page: 53 of 57]

USGS, Content Standards for Digital Geospatial Metadata, available from anonymous ftp site fgdc.er.usgs.gov/gdc/metadata/meta.6894.ps, 1994b.

Vanicek, P. and Krakiwsky, E.J. *Geodesy: The Concepts*, North-Holland, Amsterdam, 1982.

Veregin, H. (1994) "Integration of simulation modelling and error propagation for the buffer operation in GIS." *Photogrammetry Engineering and Remote Sensing*, Vol.60, No.4, pp. 427-435.

Veregin, H "A review of error models for vector to raster conversion", *Operational Geographer*, vol.1, no.1, 1989, pp.11-15.

Walpole, R.E. and Myers, R.H. *Probability and Statistics for Engineers and Scientists*, Macmillan, NY, 1972.

Wang, F. "Fuzzy supervised classification of remote sensing images", *IEEE Trans. on Geoscience and Remote Sensing*, vol.28, no.2, March 1990, pp.194-201.

WaterMark Industries Inc., "Managing uncertainty in digital geographic information: survey of technology", Phase I Report for Centre de Recherches Pour La Defense Valcartier, BFC Valcartier, Courcelette, Quebec, Feb. 16, 1995.

Williams, R.J. "Data quality statements for spatial databases", Research Note ERL-0632-RN, Defence Science and Technology Organization (DSTO) Australia, Electronics Research Laboratory, Salisbury, Australia, July, 1992.

Xu, Xiubin and Nickerson, B.G. "Applications of the extension principle to the plane", UNB Faculty of Computer Science Technical Report TR95-092, March, 1995.

Yager, R.R. "Misrepresentations and challenges A response to Elkan", *IEEE Expert*, August 1994, pp.41-42.

Zadeh, LA. "Fuzzy sets", *Information and Control*, vol.8, 1965, pp.338-353.

Zadeh, L.A. "The concept of a linguistic variable and its application to approximate reasoning. Parts 1, 2 and 3", *Information Science*, 1975, vol.8, pp.199-249, vol.8, pp.301-357, vol.9, pp.43-80.

Zadeh, L.A. "Why the success of fuzzy logic is not paradoxical", *IEEE Expert*, August 1994, pp.43-46.

Zhou, F.Q. (1995). Doctoral dissertation in preparation. Dept. of Geodesy and Geomatics Engineering, UNB.

P153167.PDF [Page: 54 of 57]

Appendix I. Digital Nautical Chart Data Quality Attributes

The Digital Nautical Chart [Department of Defence, 1993, p.189] is a DRAFT standard for representing hydrographic charts. It is designed to represent the data currently on the paper versions of Harbor, Approach, Coastal and General hydrographic charts produced by the Defense Mapping Agency. The detailed data format is Vector Product Format (VPF) which is very similar to the DIGEST Vector Relational Format (VRF). DNC has specific requirements for data quality parameters as follows:

1. SOURCE_ID Chart Identifier - the 5 digit code that uniquely identifies a DMA nautical chart.
2. NAME Chart Name - The common textual identifier of the chart, usually consisting of a hydrographic feature reference (e.g. "Hampton Roads" or "Cape May to Cape Hatteras").
3. COMP_DATE First Edition Date - The date of original chart compilation (chart edition 1).
4. EDITION Chart Edition Number - The chart edition number.
5. PRINT_DATE Chart Edition Date - The chart edition date.
6. SCALE Chart Scale - The denominator of the chart scale expressed as a fraction.
7. REF_LAT Reference Latitude - The latitude noted on the chart to which the chart scale is reference.
8. ABS_HORIZ_ACC Absolute Horizontal Accuracy - The absolute horizontal accuracy in meters.
9. ABS_VERT_ACC Absolute Vertical Accuracy - The absolute vertical accuracy in meters.
10. DATUM Hydrographic Datum - The chart hydrographic datum.
11. LINEAGE Chart Lineage - Available information as to the source of the chart data, including dates, device (sounder) used and type of positioning (e.g. GPS).
12. SOURCE_INFO General Source Information - Any pertinent information regarding the source chart not included in any other attribute.

Items included here that are not normally included for land maps includes the reference latitude, hydrographic datum, and lineage information referring to specific hydrographic instrumentation (e.g. sounders).

WaterMark

UNCLASSIFIED
SECURITY CLASSIFICATION OF FORM
(Highest classification of Title, Abstract, Keywords)

DOCUMENT CONTROL DATA

1. ORIGINATOR (name and address)

Water Mark Industries Inc.
115 Ascot Drive Fredericton, NB
Canada, E3B 6C1

2. SECURITY CLASSIFICATION
(Including special warning terms if applicable)

UNCLASSIFIED

3. TITLE (its classification should be indicated by the appropriate abbreviation (S,C,R or U)
MANAGING UNCERTAINTY IN DIGITAL GEOGRAPHICAL INFORMATION: SURVEY OF TECHNOLOGY PHASE II FINAL REPORT

4. AUTHORS (Last name, first name, middle initial. If military, show rank, e.g. Doe, Maj. John E.)

Doucette, M., Lee, Y.C., Nickerson, B., and Roberts G.

5. DATE OF PUBLICATION (month and year)	6a. NO. OF PAGES	6b. NO. OF REFERENCES
March 1995	51	71

7. DESCRIPTIVE NOTES (the category of the document, e.g. technical report, technical note or memorandum. Give the inclusive dates when a specific reporting period is covered.)

Contract Report

8. SPONSORING ACTIVITY (name and address)

Defence Research Establishment Valcartier
2459 Bd Pie XI north, P.O. Box 8800, Courcelette, Qc, G0A 1R0

9a. PROJECT OR GRANT NO. (Please specify whether project or grant)	9b. CONTRACT NO.
0312D14A	W7701-4-1824/01-XSK
10a. ORIGINATOR'S DOCUMENT NUMBER	10b. OTHER DOCUMENT NOS.
	N/A

11. DOCUMENT AVAILABILITY (any limitations on further dissemination of the document, other than those imposed by security classification)

- [X] Unlimited distribution
- [ ] Contractors in appoved countries (specify)
- [ ] Canadian contractors (with need-to-know)
- [ ] Government (with need-to-know)
- [ ] Defence departments
- [ ] Other (please specify) :

12. DOCUMENT ANNOUNCEMENT (any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in 11) is possible, a wider announcement audience may be selected.)

13. ABSTRACT (a brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is hig desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U) It is not necessary to include here abstracts in both official languages unless the text is bilingual).

The results of a survey of the state of the technology in uncertainty management techniques is presented, with specific reference to geographical data uncertainty. This survey included reviews of uncertainty data collection, modeling, representation, propagation, visualization - including animation and the application of fuzzy logic in handling spatial data. It is clear from the survey that these are active areas of research, and some of the findings can be implemented in systems to manage uncertainty of spatial data.

This study also identifies areas within the DIGEST format that might be considered to facilitate uncertainty management. The following changes to DIGEST are discussed: resolution of the internal in consistencies of DIGEST, development of a modeling language for DIGEST, addition of specifications for quality control testing, modification of the lineage data quality representation in DIGEST, enhancement of topological consistency within DIGEST, development of toolkits for generating quality assurance measures and the development of toolkits for viewing uncertainty information. These suggestions for modifications to DIGEST are cross-referenced with the issues outlined in Phase 1 of this project, giving consideration to the need to freeze DIGEST and to provide for a stable development environment.

This report concludes that with these modifications, DIGEST will be more "user friendly" and will be improved in its capability to clearly and precisely specify uncertainty in the geographical data it represents. For DIGEST to become more widely used and accepted, it must overcome the marketplace perception that it is "closed" and difficult to use.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipm model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus. e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus-identified. If it is n possible to select indexing terms which are Unclassified, the classification of each sould be indicated as with the title.)

DIGITAL GEOGRAPHICAL INFORMATION
GIS
DATA QUALITY AND UNCERTAINTY
DIGEST
FUZZY SETS